Московский государственный университет
имени М. В. Ломоносова

Научно-исследовательский институт ядерной физики
имени Д. В. Скобельцына

кафедра оптики и спектроскопии физического факультета

О. Е. НАНИЙ, А. Н. ТУРКИН

Оптические методы в информатике

Москва
Издательство «Университетская книга»
2010

УДК 535
ББК 22.34-я73-1
Н25

Наний О. Е., Туркин А. Н.
Н25 Оптические методы в информатике : Учебное пособие / О. Е. Наний, А. Н. Туркин — М. : Университетская книга, 2010. — 112 с. : табл., ил.
ISBN 978-5-91304-126-5

Данный курс лекций «Оптические методы в информатике» читается студентам кафедры оптики и спектроскопии физического факультета МГУ в 9-м семестре. С единых информационных позиций рассмотрен широкий круг оптических явлений.

Основное внимание уделено описанию физических принципов и методов, лежащих в основе применения оптики в оптической передаче и представлении информации.

УДК 535
ББК 22.34я73-1

Учебное издание

Олег Евгеньевич Наний
Андрей Николаевич Туркин
ОПТИЧЕСКИЕ МЕТОДЫ В ИНФОРМАТИКЕ

Учебное пособие

Подп. в печать 20.05.2010. Формат 60×84 $^{1}/_{16}$. Бумага офсетная.
Печать цифровая. Тираж 100 экз. Заказ № Т-095.

Отпечатано с диапозитивов, предоставленных автором, в типографии «КДУ».
Тел./факс (495) 939-44-91; www.kdu.ru; e-mail: press@kdu.ru

ISBN 978-5-91304-126-5

О Г Л А В Л Е Н И Е

Лекция 1. Информация и оптика

Понятие информации. Роль информации в жизни человека. Оптика и информация: индикация, передача, хранение и обработка информации оптическими методами. Количественные характеристики информации. Цифровые и аналоговые сигналы. Аналогово-цифровое преобразование. Пропускная способность цифрового и аналогового канала.

«Знак – это чувственно воспринимаемое в символе» (Людвиг Витгенштейн)

Понятие информации

Информация – это основное понятие информатики, уточнение содержания которого не может быть достигнуто с помощью определения, так как последнее лишь сводило бы это понятие к другим не определенным основным понятиям. Информация содержится в самых разнообразных сведениях, сообщениях, известиях, знаниях и умениях. Информация субъективна, зависит от подготовленности субъекта воспринимать информацию. В таком виде понятие информации использовать в технике связи в настоящее время невозможно.

Техническое определение информации основано на том, что при любых видах работы с информацией всегда идет речь о ее представлении в виде определенных символических структур (символов, знаков). Информация, представленная в символическом виде, является сообщением. Хотя соотношение между сообщением и информацией в широком смысле не является взаимно однозначным, для технических целей термин «информация» используется в значении содержания сообщения. С этой точки зрения, как заметил Жан Кокто, «величайшее литературное произведение - в принципе не что иное, как разбросанный в беспорядке алфавит».

В технике связи используется техническое определение: информация – это содержание сообщения. **Сообщение** – это информация, выраженная (представленная) в определенной символической форме и предназначенная для передачи от источника к пользователю (тексты, фото, речь, музыка, телевизионное изображение и др.). **Символ и знак** в науке (логике, математике и др.) – синонимы. Знаки подразделяются на языковые (входящие в некоторую знаковую систему) и неязыковые: копии (например, фотографии), признаки, символы. Создание специальных знаков, и особенно создание систем формул, обычно открывает в науке новые возможности: рационально построенные системы знаков позволяют в

обозримой форме выражать соотношения между изучаемыми явлениями; добиваться однозначности используемых терминов; фиксировать такие понятия, для которых в обычном языке нет словесных выражений.

Правило интерпретации – это некоторое отображение, связывающее сообщение N с информацией I, представляющее собой результат договоренности между отправителем и получателем сообщения или являющееся предписанием со стороны отправителя. Сообщение N, состоящее из некоторого набора знаков может быть преобразовано в другой набор знаков с помощью правила отображения, которое называется **кодом.**

При любых видах работы с информацией всегда идет речь о ее представлении в виде определенных символических структур. Наиболее распространены одномерные представления информации, при которых сообщения имеют вид последовательностей символов. Так информация представляется в письменных текстах, при передаче по каналам связи, в памяти ЭВМ. Однако широко используются и многомерные представления информации, причем под многомерностью понимают не только расположение элементов информации на плоскости или в пространстве (в виде рисунков, схем, графов, объемных макетов и т.д.), но и множественность признаков используемых символов. Например, информацию могут нести не только значения букв и цифр, но и их цвет, размер, вид шрифта.

Формирование представления информации называется ее кодированием. Часто термин "кодирование" употребляется в более узком смысле, как переход от исходного представления к представлению, удобному для хранения, передачи или обработки. В этом случае обратный переход к исходному представлению называется декодированием.

При кодировании могут ставиться разные цели и соответственно применяться разные методы. Наиболее распространенные цели кодирования - это экономность, т. е. уменьшение избыточности сообщения; повышение скорости передачи или обработки; надежность, т. е. защита от случайных искажений; сохранность, т. е. защита от нежелательного доступа к информации; удобство физической реализации (например, двоичное кодирование информации в ЭВМ); удобство восприятия. Эти цели часто противоречат друг другу. Экономные сообщения могут оказаться менее надежными, так как они не содержат лишних символов, и искажение любого символа может изменить смысл сообщения. Например, обычная запись чисел цифрами гораздо экономнее и удобнее для вычислений, чем запись

словами. Однако искажение или удаление любой цифры изменяет величину числа. Поэтому в финансовых документах, где надежность крайне важна, цифровые представления чисел в некоторых местах заменяются или дублируются словесными представлениями. (Сумму иногда пишут прописью.) Теория кодирования подробно исследует проблемы разумного сочетания экономности и надежности при передаче информации.

На разных этапах сложного информационного процесса достигаются разные цели, и поэтому информация неоднократно перекодируется, т. е. изменяет свое представление.

Сигнал – это физический процесс, имеющий информационное значение, установленное принятым соглашением. В информатике и технике связи термин «сигнал» употребляется в более широком смысле – как любой физический процесс, распространяющийся в пространстве и времени, параметры которого способны переносить сообщение. Для того, чтобы физический процесс переносил информацию хотя бы один из его параметров должен быть промодулирован в соответствии с передаваемой информацией. **Модуляция (сигнала)** – это изменение некоторой физической величины (параметра сигнала) во времени, обеспечивающее передачу сообщения. Сигналы могут быть дискретными и аналоговыми.

Дискретный сигнал – это сигнал, параметр которого может принимать лишь конечное число значений. Сообщение, представленное в виде конечного числа символов, и которое следовательно может быть передано с помощью дискретных сигналов называется **дискретным сообщением.**

Роль информации в жизни человека

Информационные процессы, т. е, процессы передачи, хранения и переработки информации, всегда играли важную роль в жизни общества. Люди обмениваются устными сообщениями, записками, посланиями. Они передают друг другу просьбы, приказы, отчеты о проделанной работе, описи имущества; публикуют рекламные объявления и научные статьи; хранят старые письма и документы; долго размышляют над полученными известиями или немедленно кидаются выполнять указания начальства. Все это - информационные процессы.

Информация всегда связана с материальным носителем, а ее передача - с затратами энергии. Однако одну и ту же информацию можно хранить в различном материальном виде (на бумаге, в виде фотонегатива, на магнитной ленте) и передавать с различными

энергетическими затратами (по почте, по телефону, с курьером и т. д.), причем последст вия - в том числе и материальные - переданной информации совершенно не зависят от физических затрат на ее передачу. Например, легкое нажатие кнопки опускает тяжелый театральный занавес или взрывает большое здание, красный свет светофора останавливает поезд, а неожиданное неприятное известие может вызвать инфаркт. Поэтому информационные процессы не сводимы к физическим, и информация, наряду с материей и энергией, является одной из фундаментальных сущностей окружающего нас мира.

Достижения техники в 18 - 19 вв. практически целиком были связаны с успехами физики и химии. Благодаря им были созданы и широко распространились различные преобразователи материи и энергии: двигатели, металлургические и химические производства, электрогенераторы и т. д. Эффективность их работы описывается с помощью физических понятий: мощности, к. п. д., грузоподъемности, количества вырабатываемой энергии и др. В 20 в. с развитием техники появились устройства другого рода: средства связи, устройства автоматики, а с 40-х гг. - вычислительной техники. Выяснилось, что эффективность их работы с помощью физических понятий описать невозможно и что существенные характеристики таких устройств нужно описывать совсем другими способами. В результате впервые возникли точное понятие информации и математическая теория информации. Стало ясно, что средства связи, какие бы физические процессы они ни использовали, - это средства передачи информации. Объединение понятий "информация" и "управление" привело Н. Винера в 40-х гг. к созданию кибернетики* которая, в частности , впервые указала на общность информационных процессов в технике, обществе и живых организмах. Использование понятия информации оказало существенное влияние на развитие современной биологии, особенно таких ее разделов, как нейрофизиология и генетика. И наконец, в связи с развитием вычислительной техники, стимулировавшей информатизацию всего общества, возник комплекс наук о различных аспектах работы с информацией - информатика.

Оптика и информация: индикация, передача, хранение и обработка информации оптическими методами

Оптические методы играли в информатике одну из ключевых ролей на протяжении развития всей истории человечества. Человеческий глаз – главнейший канал поступления информации для

человека и оптикой первоначально называлась наука о формировании изображений в видимом свете и восприятии изображений человеком. Первые системы коммуникации между людьми были оптическими (жесты, мимика), древние системы передачи информации на большие расстояния также использовали свет (сигнальные костры, факелы, оптический телеграф). Возможность закрепления знаний в обществе и ее хранение отдельно от индивидуумов появилась тогда, когда были созданы первые изображения. Возникновение письменности а затем книгопечатания сыграли революционную роль в развитии информационной среды человеческого общества. Существенное увеличение информационной емкости хранимой информации было достигнуто при изобретении фотографии.

По современным представлениям классической физики свет представляет собой электромагнитные волны определенного диапазона длин волн или частот, называемого оптическим. Границы оптического диапазона определяются условно примерно от долей миллиметра до десятков нанометров (границы этого диапазона по частоте: от десятков ГГц до ТГц). Таким образом, видимый свет – это узкая спектральная область на шкале электро-магнитных волн (см. рис.1). Световые волны отличаются от радиоволн в принципе только одним – частотой (в оптике вместо частоты ν чаще используется длина волны $\lambda=c/\nu$, c – скорость света в вакууме). Однако по своим физическим свойствам и характеру распространения в пространстве световые и радиоволны существенно различаются, из чего вытекают технологические отличия ВОЛС от традиционных систем электросвязи.

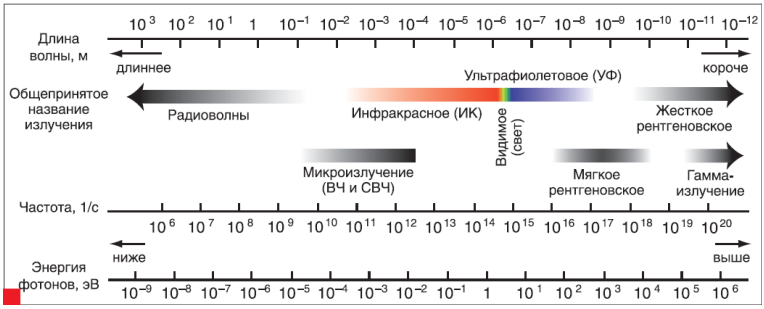

Рис.1.1. Спектр электромагнитных волн

В настоящее время оптические методы используются в той или иной степени на всех стадиях преобразования информации:

- При получении информации
- Передаче информации
- Хранении информации
- Обработке информации
- Представлении информации (индикации).

В наибольшей степени оптические методы используются в системах индикации и передачи информации. В системах связи – оптика вытесняет электромагнитные волны других диапазонов. Среди нескольких причин две важнейшие: высокая частота несущей и малая длина волны. Основной диапазон – ближний инфракрасный.

По мере развития оптических систем связи возрастает интерес к оптическим методам обработки информации. И хотя надежды на разработку оптических компьютеров (процессоров) сегодня не очень радужны, предпринимаются значительные усилия по созданию оптических регенераторов, оптических коммутаторов и других типов устройств обработки оптических сигналов в высокоскоростных сетях связи.

В системах хранения информации – успехи оптики заметны, но пока она играет вспомогательную роль.

Количественные характеристики информации

Математическое понятие информации связано с ее измерением. В теории информации принят энтропийный подход, который учитывает ценность информации, содержащейся в сообщении для его получателя, и исходит из следующей модели. Получатель сообщения имеет определенные представления о возможных наступлениях некоторых событий. Эти представления в общем случае недостоверны и выражаются вероятностями, с которыми он ожидает то или иное событие. Общая мера неопределенности (энтропия) характеризуется некоторой математической зависимостью от совокупности этих вероятностей. Количество информации в сообщении определяется тем, насколько уменьшается эта мера после получения сообщения. Например, тривиальное сообщение, т. е. сообщение о том, что получателю и без того известно, не изменяет ожидаемых вероятностей и не несет для него никакой информации. Сообщение несет полную информацию о данном множестве событий, если оно целиком снимает всю неопределенность. В этом случае количество информации в нем равно исходной энтропии.

В технике часто используют более простой и грубый способ измерения информации, который можно назвать объемным. Он основан на подсчете числа символов в сообщении, т. е. связан с его длиной и не учитывает содержания. Правда, длина сообщения зависит от числа различных символов, употребляемых для записи сообщения, т. е. от мощности алфавита. Например, одно и то же число "девятнадцать" в десятичном алфавите записывается двумя символами - 19, а в двоичном алфавите -пятью символами - 10111. В вычислительной технике применяются две стандартные единицы измерения: бит и байт. Бит - это один символ двоичного алфавита. Байт - это один символ, который можно представить восьмиразрядным двоичным кодом; мощность алфавита этого представления равна числу различных восьмиразрядных двоичных кодов, т. е. 256, и может включать, например, все символы клавиатуры пишущей машинки или компьютера.

Эти два способа измерения информации, как правило, не совпадают, причем энтропийное количество информации не может быть больше числа двоичных символов (битов) в сообщении. Если же оно меньше этого числа, то говорят, что сообщение избыточно. Тривиальные сообщения всегда избыточны, так как имеют нулевую информацию с точки зрения энтропии, но содержат ненулевое число символов.

Мы будем пользоваться **технической мерой количества информации** как мерой «затрат», необходимых для классификации знаков. В соответствии с этим количество информации в знаке

$$h_i = \log_2\left(\frac{1}{p_i}\right), \qquad (1.1)$$

где p_i - вероятность появления знака. Можно определить среднюю информацию на знак в системе знаков:

$$H = \sum p_i \log_2\left(\frac{1}{p_i}\right) \le \log_2 n \qquad (1.2)$$

Неравенство(1.2) показывает, что наиболее рационально использовать все знаки с равной вероятностью. Если количество знаков есть степень 2, т.е. $n = 2^N$ и знаки равновероятны ($p_i = (1/2)^N$) то $H = \log_2(n) = N$.

Цифровые и аналоговые сигналы.

В информатике и технике связи **сигналом** называется любой процесс, несущий информацию. Он может иметь произвольную физическую природу: механическую (движение, давление), тепловую,

световую, электрическую, акустическую. Среди параметров процесса выбирается один или несколько, значения которых должны нести информацию. Такими параметрами могут быть длительность, амплитуда, частота, яркость, цвет и т. д. Необязательно, чтобы физически различные значения выбранного параметра соответствовали различной информации. Например, в сигнализации на транспорте информационный смысл имеют три значения цвета светофора: красный, желтый и зеленый. Разные оттенки красного или зеленого не играют роли.

Если выбранные информационные значения образуют дискретное множеством сигнал называется дискретным, или цифровым. Если это множество непрерывно, сигнал называется непрерывным, или аналоговым. Светофор - пример дискретного сигнала. Передача сообщений с помощью азбуки Морзе также дискретна. Здесь информационным параметром является длительность сигнала, а возможные значения этого параметра - короткие и длинные сигналы, а также короткие и длинные паузы между ними. Сигналы, несущие информацию об изменениях температуры, напряжения, давления и других физических характеристик, обычно непрерывны. Сигналы, несущие текстовую, символическую информацию - дискретны.

Различная аппаратура систем обработки информации, вычислительной техники и управления в зависимости от того, какие сигналы она обрабатывает, делится на аналоговую и дискретную. В системах передачи и обработки информации сигналы обычно неоднократно преобразуются. При этом их физическая природа может меняться без потери информации. Может измениться и информационный характер сигнала: аналоговый сигнал может быть преобразован в дискретный (такой процесс называется аналогово-цифровым преобразованием или дискретизацией) и наоборот. Например, при вводе результатов измерения непрерывных величин в компьютер происходит дискретизация сигналов.

Аналогово-цифровое преобразование

Первым шагом в преобразовании аналогового сигнала в цифровой сигнал является определение значений сигнала (отсчетов) через одинаковые интервалы времени, как это показано на рис.1.2. Этот процесс называется дискретизацией по времени. В соответствии с теоремой Котельникова (теоремой отсчетов) для точного представления аналогового сигнала (рис. 1.2, а) дискретным (рис. 1.2, *в)* необходимо, чтобы частота отсчетов (дискретизации) f_s, равная 1/T, где T — интервал дискретизации, была, по крайней мере, в 2 раза

больше наивысшей частоты f_m, содержащейся в спектре дискретизируемого сигнала. При выполнении этого условия исходный аналоговый сигнал может быть просто восстановлен путем пропускания дискретизированного сигнала через фильтр низких частот, пропускающий все частоты ниже f_m. Диапазон частот от 0 до f_m представляет собой ширину спектра исходного сигнала, которую будем обозначать Δf. Таким образом, частота дискретизации должна выбираться из условия $f_s > 2\Delta f$. Следующий шаг – дискретизация по амплитуде. Он заключается в том, что отсчетам (т.е. значениям сигнала) присваиваются значения из определенного конечного множества значений. В примере на рис. 1.2 такими значениями являются целые числа, это означает, что шаг дискретизации (квантования) по амплитуде равен 1. Общее число уровней квантования определяется делением диапазона изменения исходного непрерывного сигнала на шаг квантования.

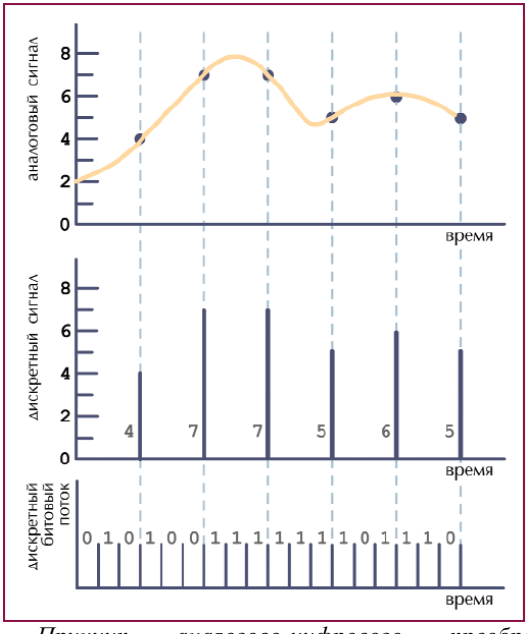

Рис.1.2. Принцип аналогово-цифрового преобразования: а) исходный аналоговый сигнал; б) дискретный сигнал; в) дискретный битовый поток

Пропускная способность цифрового и аналогового канала

Пропускная способность B цифрового канала, очевидно, определяется произведением количества переданных в секунду символов K_S на среднее количество информации, приходящееся на один символ. Если информация кодируется оптимально, то среднее количество информации, приходящееся на один символ определяется выражением $\log_2 m$, где m - количество значений (уровней), которые может принимать передаваемый символ. В этом случае пропускная способность канала определяется следующим выражением:

$$B = K_S \cdot \log_2 m = \frac{1}{t_S} \log_2 m \qquad (1.3)$$

В двухуровневых системах связи $B = K_S$. Переход к многоуровневым системам связи позволяет увеличить пропускную способность канала не увеличивая скорость передачи символов.

Оценим **информационную пропускную способность аналогового канала**, который способен передавать аналоговый сигнал, занимающий полосу частот Δf, и поддерживать на выходе приемника (где отношение сигнал-шум наименьшее) отношение пикового значения мощности сигнала к среднеквадратичному значению мощности шума, равное (P_S / P_N).

Воспользуемся теоремой отсчетов, в соответствии с которой для точного представления переданного аналогового сигнала цифровым, через интервалы времени $t_S = 1/2\Delta f$ нужно передавать цифровое сообщение, а именно амплитудное значение сигнала. Пусть число уровней квантования равно m. Тогда каждый отсчет сигнала потребует для своего кодирования $N = \log_2 m$ двоичных цифр. Увеличивая число уровней квантования можно увеличивать поток передаваемой информации. Однако наличие шумов ограничивает этот процесс. Отношение пиковой мощности сигнала (P_S) к среднеквадратическому значению мощности шума (P_N) определяет максимальное число уровней квантования M:

$$\mathrm{M} = [1 + (P_S / P_N)]^{1/2}. \qquad (1.4)$$

Таким образом, исходный аналоговый сигнал, занимающий полосу частот Δf (Гц) и имеющий динамический диапазон A_S / A_N можно передать в цифровом виде используя B двоичных цифр в секунду (бит/с), где

$$B = 2\Delta f \log_2 [1 + (P_S / P_N)]^{1/2} = \Delta f \log_2 [1 + (P_S / P_N)] \qquad (1.5)$$

Формула (1.5),определяющая величину пропускной способности *B* канала, получила широкую известность как формула Шеннона.

Лекция 2. Элементы зонной теории полупроводников

Функции Блоха и уровни энергии. Распределение электронов по уровням энергии. Поглощение света полупроводниками при межзонных переходах. Спонтанное излучение полупроводников.

Функции Блоха и уровни энергии

Волновая функция электрона в идеальном кристалле, обладающем трансляционной симметрией, может быть записана в виде

$$\psi_k(r) = \exp(i\vec{k}\cdot\vec{r})u_k(\vec{r}), \qquad (2.1)$$

где $u_k(\vec{r})$ - периодические функции с периодом кристаллической решетки. Волновые функции такого вида называют *функциями Блоха*, а постоянная распространения $\vec{k}$ может служить "квантовым числом", характеризующим состояние, описываемое такой волновой функцией.

Можно ввести вектор квазиимпульса $\vec{P}$, связанный с постоянной распространения (волновым вектором) $\vec{k}$, соотношением:

$$\vec{P} = \hbar\vec{k} \qquad (2.2)$$

Квазиимпульс $\vec{P}$ является интегралом движения (при движении электрона, находящегося в некотором состоянии (2.1), он остается постоянным), и так же, как и $\vec{k}$, его можно использовать в качестве "квантового числа".

Поскольку на границах кристалла должны выполняться симметричные граничные условия, то постоянная распространения $\vec{k}$ квантуется следующим образом:

$$k_i = \frac{2\pi s}{L_i} \qquad (2.3)$$

где $i = x, y, z$; s – целое число; L_i – длина кристалла в i-том направлении. Объем в $\vec{k}$-пространстве, приходящийся на одно состояние электрона с данным значением $\vec{k}_i$, равно $8\pi^2/V$, где $V = L_x L_y L_z$. В состоянии с данным значением $\vec{k}_i$ могут находиться в соответствии с принципом Паули только два электрона с разными значениями спина. Следовательно полное число разрешенных электронных состояний, соответствующих значениям k в интервале от k до $k + dk$ равняется удвоенному объему шарового слоя с радиусом k и толщиной dk, деленному на объем, приходящийся на одно электронное состояние, т.е.

$$\rho(k)dk = \frac{k^2 V}{\pi^2} dk \,. \qquad (2.4)$$

В общем случае зависимость энергии электронов в полупроводниках от $\vec{k}$ может быть очень сложной. Однако вблизи дна зоны проводимости энергию электрона E можно считать зависящей только от значения модуля волнового вектора $k = |\vec{k}|$, причем эта зависимость имеет вид:

$$E(k) = E_C + \frac{\hbar^2 k^2}{2m_e^*}; \qquad (2.5)$$

где m_e^* – постоянная, имеющая размерность массы. Ее значение не совпадает с массой свободного электрона $(m_e^* \neq m_e)$. Энергетические зоны вида (2.5) обычно называют “сферическими”, т.к. поверхности равной энергии в $\vec{k}$ - пространстве имеют в этом случае форму сфер. Зоны такой формы следует считать простейшей приближенной моделью реальных, значительно более сложных зон.

Аналогичным образом простейшей моделью вершины валентной зоны также является сферическая зона. Если максимум энергии $E(k)$ достигается в точке $k = 0$, то зависимость $E(k)$ имеет вид:

$$E(k) = E_V - \frac{\hbar^2 k^2}{2m_{eV}^*}. \qquad (2.6)$$

где m_{eV}^* – постоянная, имеющая размерность массы, но в общем случае отличающаяся от m_e^*, дополнительный индекс V означает, что этот коэффициент пропорциональности относится к валентной зоне.

Поскольку валентная зона почти полностью заполнена электронами, и в ней имеется лишь незначительное количество незаполненных (вакантных) состояний, то такие вакантные состояния удобно описывать как квазичастицы, получившие название дырок. Дыркам следует приписать положительный заряд $+e$ и эффективную массу $m_h^* = m_{eV}^*$.

Таблица 2.1. Эффективные массы электронов и дырок некоторых полупроводников

Материал	GaAs	InP	GaN	GaP	Si	Ge
$\lvert m_e^* \rvert / m_e$	0,067	0,08	0,20	0,82	0,98	1,64
$\lvert m_h^* \rvert / m_e$	0,45	0,56	0,80	0,60	0,49	0,28

На рис. 2.1 показана типичная зависимость $E(k)$ для полупроводника с прямым переходом, т.е. для полупроводника, в котором минимуму зоны проводимости и максимуму валентной зоны соответствует одно и то же значение волнового вектора $\vec{k}$. На рисунке показан случай $m_h^* > m_e^*$, хотя возможна и противоположная ситуация.

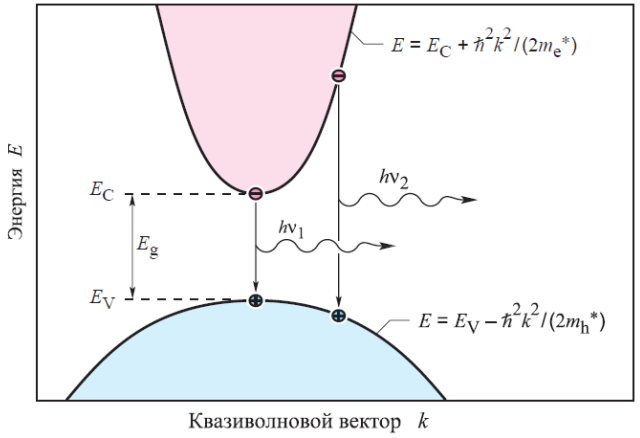

Рис.2.1. Уровни энергии в зоне проводимости и валентной зоне прямозонных полупроводников. Показаны также излучательные переходы электронов из зоны проводимости в валентную зону. Из закона сохранения квазиимпульса следует, что при таком излучательном переходе квазиимпульс и волновой вектор электрона сохраняются.

Из выражения (2.4) для плотности состояний в $\vec{k}$-пространстве можно получить выражения для плотностей состояний в энергетическом пространстве в единице объема кристалла:

$$\rho_C(E) = \frac{1}{V}\rho(k)dk\frac{dk}{dE} = \frac{1}{2\pi^2}\left(\frac{2m_C^*}{\hbar^2}\right)^{3/2}(E - E_C)^{1/2} \qquad (2.7)$$

$$\rho_V(E) = \frac{1}{V}\rho(k)dk\frac{dk}{dE} = \frac{1}{2\pi^2}\left(\frac{2m_h^*}{\hbar^2}\right)^{3/2}(E_h - E)^{1/2} \qquad (2.8)$$

где индексы C и V соответствуют зоне проводимости и валентной зоне. В формулах (2.7) и (2.8) использована взаимосвязь между k и E, следующая из формул (2.5) и (2.6).

Распределение электронов по уровням энергии

В теории твердого тела объяснение ряда фундаментальных физических процессов и свойств основывается на модели *свободного электронного газа Ферми* - системы (совокупности) свободных невзаимодействующих электронов, подчиняющихся принципу Паули. В соответствии с такой моделью вероятность того, что состояние с энергией E занято электроном, дается распределением Ферми-Дирака $P_e(E,T)$ вида:

$$P_e(E,T) = \frac{1}{\exp[(E-E_F)/k_BT]+1}, \qquad (2.9)$$

где E_F – энергия Ферми, T – температура, k_B – постоянная Больцмана

Если абсолютная температура равна нулю (T =0), то функция $P_e(E,T)$ имеет вид ступенчатой функции: все уровни, энергия которых меньше E_F заняты электронами с вероятностью 1 ($P_e(E,T)$ =1), а все уровни с энергией больше энергии Ферми не заняты ($P_e(E,T)$ =0).

В металлах уровень Ферми попадает в разрешенную зону, которая, таким образом, оказывается частично заполненной электронами.

В *диэлектриках или полупроводниках* при T =0 последняя из зон, которая еще содержит электроны, целиком заполнена ими (эта зона называется валентной), а расположенная над ней зона пуста (эта зона называется зоной проводимости). Различие между диэлектриками и полупроводниками не носит принципиального характера и заключается лишь в величине ширины запрещенной зоны, отделяющей край (дно) зоны проводимости E_c от края (потолка) валентной зоны E_v : .

Энергия запрещенной зоны, или *ширина энергетической щели* E_g равна разности энергий E_c и E_v :

$$E_g = E_c - E_v. \qquad (2.10)$$

К диэлектрикам обычно относят вещества с E_g > 3 эB. Так, например, типичными диэлектриками являются алмаз и корунд, у которых величина E_g превышает 5 эB. Значения величины запрещенной зоны для некоторых полупроводников приведены в таблице 2.2.

Таблица 2.2. Запрещенные зоны некоторых полупроводников

Материал	GaAs	InP	GaN	GaP	Si	Ge
E_g , эВ	1,42	1,35	3,4	2,26	1,12	0,66

В термодинамически равновесной системе значение уровня Ферми в любой ее части одинаково. В неравновесной системе уровни Ферми валентной зоны и зоны проводимости могут отличаться.

На рис. 2.2 изображены схемы энергетических зон в полупроводниках с проводимостью различных типов. Зоны, приведенные на схемах, имеют простейшую, "сферическую" форму. Из вида зависимостей E от k следует, что $|m_h^*| = |m_e^*|$. Черными точками на схемах условно (не в масштабе) показаны уровни энергии, занятые электронами. "Белые" точки (кружки) соответствуют "пустым" уровням, т.е. уровням, не занятым электронами, или, говоря иначе, "занятыми дырками".

Схема *а*) на рис. 2.2 иллюстрирует случай собственного (нелегированного) полупроводника при $T = 0$. Уровень Ферми расположен посередине запрещенной зоны и ни с каким реальным уровнем энергии в кристалле не совпадает.

При легировании полупроводников примесями разного типа можно сдвигать положение уровня Ферми. Схемы *б*) и *в*) иллюстрируют случаи сильно легированных полупроводников (примесные уровни на схемах не показаны). В первом случае уровень Ферми E_F оказывается расположенным в зоне проводимости, во втором - в валентной зоне. Примесные атомы, которые с достаточно большой вероятностью могут отдавать свои электроны в зону проводимости, называют *донорными*, или *примесями n-типа*.

Примесные атомы, которым энергетически выгодно "выхватить" электрон из валентной зоны, называются *акцепторными*, или *примесями p-типа*.

Полупроводники, у которых уровень Ферми расположен внутри зоны проводимости или валентной зоны, называют *вырожденными*. На рис. 2.2 схема *б*) соответствует *вырожденному полупроводнику n-типа*, а схема *в*) - *вырожденному полупроводнику p-типа*. Они обладают соответствующими типами проводимостей.

В термодинамически неравновесном полупроводнике может возникнуть и такое распределение носителей, которое в зоне проводимости имеет вид, характерный для равновесного вырожденного полупроводника n-типа, а в валентной зоне - для полупроводника p-типа. Такую область неравновесного полупроводника называют дважды вырожденной (схема г). Она характеризуется двумя квазиуровнями Ферми: уровнем E_F^c в зоне проводимости и уровнем E_F^v в валентной зоне.

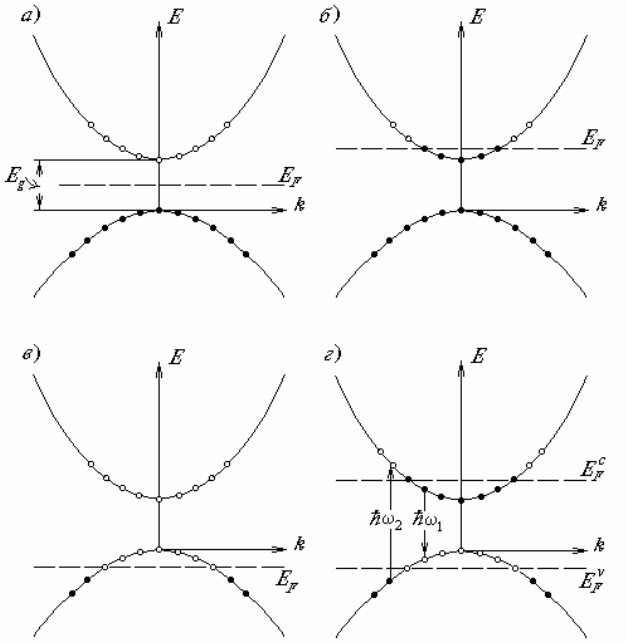

Рис.2.2. Энергетические зоны полупроводников разного типа при T = 0: *а) Собственный полупроводник; б) Вырожденный полупроводник n-типа; в) Вырожденный полупроводник p-типа; г) Дважды вырожденный полупроводник.*

Неравновесное состояние может быть создано, например, с помощью *оптической накачки*. Частота накачки ω_2 должна в этом случае удовлетворять условию:

$$\omega_2 > \frac{E_g}{\hbar}. \qquad (2.11)$$

Поглотившие квант $\hbar\omega_2$ электроны будут переходить из валентной зоны в зону проводимости. Электроны, попавшие в зону проводимости, и дырки, образовавшиеся в валентной зоне, будут стремиться "спуститься" ко дну или "подняться" к вершине соответствующих зон с характерным для этих процессов временем релаксации $\sim 10^{-13}$ *с*. Положения верхней границы состояний, заполненных электронами в зоне проводимости (квазиуровень

Ферми E_F^c) и нижней границы дырок в валентной зоне (E_F^v) зависят от интенсивности накачки, и для возникновения этих квазиуровней она должна быть достаточно сильной. С ее ростом разность E_F^c - E_F^v увеличивается. Другой способ создания дважды вырожденных полупроводников – инжекция неосновных носителей, который будет рассмотрен ниже.

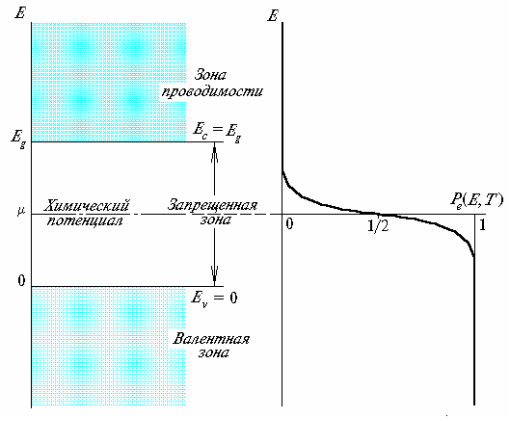

Рис.2.3. *Схема энергетических зон полупроводника с проводимостью собственного типа. Эффективные массы носителей одинаковы. Справа показан вид функции распределения Ферми-Дирака*

На рис. 2.3 приведена схема энергетических зон полупроводника с проводимостью собственного типа. В случае, который изображен на рисунке, абсолютные величины эффективных масс обоих видов носителей одинаковы ($|m_e^*| = |m_h^*|$), уровень Ферми (он же химический потенциал) расположен точно посередине между валентной зоной и зоной проводимости, и его положение при изменении температуры не меняется. Если $|m_e^*| \neq |m_h^*|$, то при изменении температуры положение химического потенциала (или “уровня Ферми”) будет меняться. Однако в большинстве собственных полупроводников при обычных температурах отклонение его положения от середины запрещенной зоны невелико.

Поглощение света полупроводниками при межзонных переходах

В рамках теории возмущений квантовой механики можно рассчитать оптическую восприимчивость полупроводника межзонных переходах а также вероятности переходов с поглощением фотонов и

вероятности вынужденного испускания. При прямых межзонных переходах коэффициент поглощения α имеет вид:

$$\alpha(\omega) = \frac{e^2 h}{4\pi^2 m_e^2 \varepsilon_0 c n \hbar\omega} \left| H'_{AB} \right|^2 \frac{(2m_r^*)^{3/2}}{\hbar^3} (\hbar\omega - E_g)^{1/2} [(P_{eh}(E_B, T) - P_{ec}(E_A, T)], \quad (2.13)$$

где $P_{ec}(E,T) = \frac{1}{\exp[(E - E_F^c)/k_B T] + 1}$, $P_{eh}(E,T) = \frac{1}{\exp[(E - E_F^h)/k_B T] + 1}$ (функции Ферми), H'_{AB} матричный элемент оператора возмущения под действием электрического поля частоты ω,

При нормальных условиях (термодинамическое равновесие и комнатная температура) и при более низких температурах $P_{eh}(E,T) \cong 1$, а $P_{ec}(E,T) \cong 0$. В этом случае при численных расчетах формулу (2.13) записывают в виде

$$\alpha_0(\omega) = K_0 (\hbar\omega - E_g)^{1/2} \quad (2.14)$$

где коэффициент K_0 определяется экспериментально. Формула (2.14) справедлива для прямозонного полупроводника с собственной проводимостью при температуре, удовлетворяющей условию $T << E_g / k_B$. Такая ситуация изображена на рис. 2.2 а. Показатель поглощения дважды вырожденного полупроводника, изображенного на рис. 3.2. г., в области $E_g < \hbar\omega < (E_F^c - E_F^h)$ изменяет знак, т.е. наблюдается усиление.

Физический смысл формулы (2.14) легко понять, выразив энергию фотона через приведенную эффективную массу электрона:

$$\hbar\omega = \frac{\hbar^2 k^2}{2m_r^*} + E_g \quad (3.15)$$

и определив из (2.15) комбинированную плотность состояний

$$\rho(\omega) = \frac{1}{2\pi^2} \left(\frac{2m_r^*}{\hbar^2} \right) (\hbar\omega - E_g)^{1/2} \quad (2.16)$$

Сравнив (2.14) и (2.16) легко заметить, что коэффициент поглощения пропорционален комбинированной плотности состояний.

Спонтанное излучение полупроводников

Оптические параметры светодиодов напрямую связаны с процессами спонтанной излучательной рекомбинации. Такие процессы схематически показаны на рис. 2.1.

Используя принцип детального равновесия можно показать, что в объемном полупроводнике процессы спонтанного излучения, вынужденного излучения и поглощения взаимосвязаны. Зная комбинированную плотности состояний $\rho(\omega)$ (2.16) и законы

распределения носителей в зоне проводимости и валентной зоне нетрудно определить спектральную зависимость интенсивности излучения

$$I(E) \sim \sqrt{E - E_g} \cdot f_c^e(E_C) f_v^h(E_V). \qquad (2.17)$$

Распределение электронов в зоне проводимости невырожденного полупроводника определяется функцией Ферми $f_c^e(E_C) = \dfrac{1}{\exp[(E_C - E_F)/k_B T] + 1}$, а распределение дырок в валентной зоне $f_v^h(E_V) = 1 - \dfrac{1}{\exp[(E_V - E_F)/k_B T] + 1} = \dfrac{1}{\exp[(E_F - E_V)/k_B T] + 1}$. При нормальных условиях обе эти функции аппроксимируются распределением Больцмана:

$$f_B(E) = \exp(-E/kT). \qquad (2.18)$$

Зависимость *интенсивности излучения* от энергии является функцией, пропорциональной произведению уравнений (2.16) и (2.18):

$$I(E) \sim \sqrt{E - E_g} \cdot \exp(-E/kT). \qquad (2.19)$$

На рис.2.4 показан соответствующий спектр излучения светодиодов.

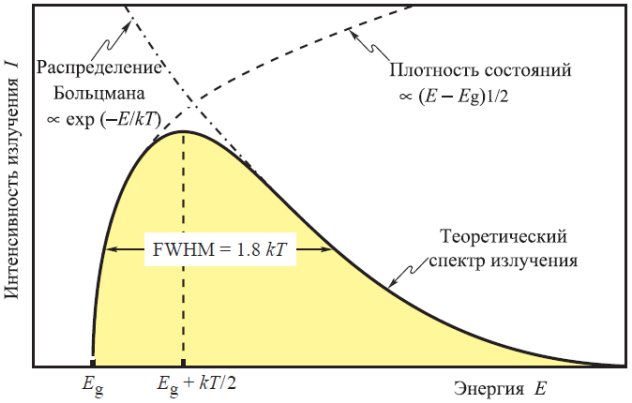

Рис.2.4. Теоретический спектр излучения светодиодов. Полная ширина спектра излучения на уровне половины максимума (FWHM) равна 1,8 kT.

Максимум спектра излучения и ширина спектральной линии соответственно равны $E = E_g + \dfrac{1}{2}kT$ и $\Delta E = 1.8kT$ ($\Delta\lambda = \dfrac{1.8kT \cdot \lambda^2}{hc}$).

Лекция 3. Полупроводниковые светодиоды

Электролюминесценция. Распределение носителей в гомогенных p-n переходах. Распределение носителей в двойных гетероструктурах. История создания полупроводниковых светодиодов.

Электролюминесценция

Открытое в начале XX века явление электролюминесценции, заключающееся в излучении фотонов твердым телом под воздействием электрического тока, позволило создать источники излучения, работающие при комнатной температуре. Суть явления заключается в том, что в некоторой области полупроводника (излучающая область) создается неравновесное распределение электронов и дырок за счет их инжекции в эту область при прохождении электрического тока. Поскольку в излучающую область должны поступать и электроны и дырки, она должна располагаться между областями с электронной и дырочной проводимостью.

Простейшая светоизлучающая структура – *p-n*-переход, схематическое изображение которого приведено на рис. 3.1. В отсутствии внешнего электрического поля в области перехода возникает потенциальный барьер, препятствующий проникновению неосновных носителей тока в смежные области.

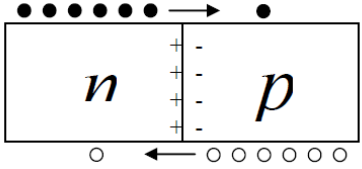

Рис. 3.1. Схематическое изображение p-n перехода.

Если приложить к полупроводниковому кристаллу внешнее напряжение так, что положительный потенциал источника приложен к *p*-области, а отрицательный - к *n*-области, то направление внешнего поля оказывается противоположным направлению поля, образованного слоями зарядов, и высота потенциального барьера уменьшается. Такое изменение барьера называют прямым смещением.

При прямом смещении число основных носителей, преодолевающих барьер, возрастает. Попав из одной области в другую, носители, как уже было замечено выше, из основных для прежней области превращаются в неосновные для новой, проникая в нее на глубину, которая определяется рекомбинационными

процессами. **Этот процесс называют инжекцией.** Инжекция носителей через барьер приводит к увеличению концентрации неосновных носителей как в *p*-, так и в *n*-области. Одновременно через контакты, к которым приложено внешнее напряжение, в эти области полупроводника поступает такое же количество основных (для каждой области) носителей. Они компенсируют излишний заряд, который вносится в каждую область инжектированными через *p*-*n*-переход неосновными носителями. Таким образом, приложение внешнего напряжения вызывает прохождение через полупроводник и, в частности, через *p*-*n*-переход тока инжекции I_{inj} .

Действие полупроводниковых излучающих кристаллов представлено схемами, приведенными на рис. 3.2. Если к полупроводниковому *p*-*n*-переходу не приложено никакого внешнего напряжения (“нулевое смещение”) и обе его части находятся в состоянии термодинамического равновесия друг с другом (схема *а*), то значение энергии уровня Ферми (химического потенциала) на всем протяжении кристалла одинаково. На схеме *а*) этому соответствует один и тот же (условный) верхний уровень заполнения состояний электронами в *p*- и *n*-областях.

Схема *б*) иллюстрирует изменение взаимного расположения энергетических зон и их заполнения электронами при прямом смещении, когда к активному кристаллу приложено внешнее напряжение V_{np}. Величина этого напряжения приблизительно равна энергетической ширине запрещенной зоны кристалла, т.е.:

$$V_{np} \approx \frac{E_g}{e}, \qquad (3.4)$$

где *e* - абсолютная величина заряда электрона.

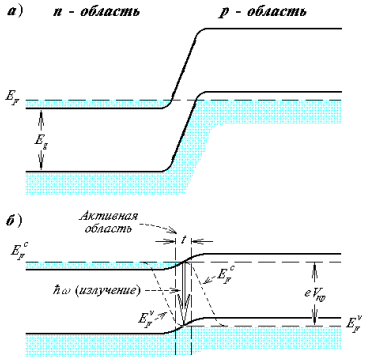

Рис.3.2. Зонные схемы, поясняющие возникновение активной области в полупроводнике с p-n-переходом. Заштрихованы области энергетических зон, заполненные электронами. а) Внешнее напряжение равно нулю. б) Внешнее напряжение $V_{np} \approx E_g / e$ *(прямое смещение).*

При приложении внешнего напряжения величина потенциального барьера уменьшается, термодинамическое равновесие между p- и n-областями кристалла нарушается и положения уровней (квазиуровней) Ферми по обе стороны p-n-перехода оказываются теперь неодинаковыми. В n-области этот уровень (его энергия обозначена через E_F^c) расположен в зоне проводимости, а в *p*-области (здесь его энергия обозначена через E_F^v) - в валентной зоне.

В случае, когда выполнено условие (3.4), в области потенциального барьера появляется область, которая содержит электроны в зоне проводимости и дырки в валентной зоне. Именно в этой области происходит рекомбинации электронов и дырок, с последующим излучение кванта света, т.е. излучательная рекомбинация. Спектр излучения при спонтанных переходах определяется формулой (2.19) с максимумом, определяемым (2.20) и шириной спектра, определяемой (2.21).

Распределение носителей в гомогенных p-n переходах

Распределение носителей тока в гомогенных p-n переходах, т.е. в переходах в пределах одного материала, зависит от коэффициента диффузии носителей. *Коэффициент диффузии* носителей измерить достаточно трудно. Гораздо проще экспериментально определить *подвижность носителей*, используя для этого, например, эффект Холла, а коэффициент диффузии получить из ***соотношения Эйнштейна***, которое для невырожденных полупроводников имеет вид:

$$D_n = \frac{kT}{e}\mu_n \text{ и } D_p = \frac{kT}{e}\mu_p. \qquad (3.5)$$

Носители, инжектированные в нейтральный полупроводник в отсутствие внешних электрических полей, перемещаются за счет диффузии. При инжектировании носителей в область с проводимостью противоположного типа, неосновные носители начинают рекомбинировать случайным образом. Среднее расстояние, которое пролетают неосновные носители до рекомбинации, называется ***диффузионной длиной***. Электроны, инжектируемые в область p-типа, до рекомбинации с дырками в среднем диффундируют на расстояние, равное диффузионной длине L_n. Для нахождения диффузионной длины используют выражения:

$$L_n = \sqrt{D_n \tau_n}; \quad L_p = \sqrt{D_p \tau_p}\ , \qquad (3.6)$$

где τ_n и τ_p – времена жизни неосновных носителей: электронов или дырок. В типичных полупроводниках диффузионная длина равняется

нескольким микронам. Например, диффузионная длина электронов в GaAs p-типа определяется как $L_n \approx 15$мкм. Таким образом, неосновные носители способны диффундировать на достаточно большое расстояние, при этом снижается их концентрация. Следовательно, зона рекомбинации расширяется становится сильно неоднородной по концентрации неосновных носителей. Такое расширение области рекомбинации в гомогенных переходах отрицательно сказывается на эффективности.

Распределение носителей в двойных гетероструктурах

Практически все современные светодиоды изготавливаются на основе двойных гетероструктур. Гетероструктуры или гетеропереходы состоят из полупроводников двух типов: с узкой запрещенной зоной для создания активной области и с широкой запрещенной зоной – для формирования барьерных слоев. Если в состав структуры входят два барьерных слоя, она называется ***двойной гетероструктурой*** (часто используется сокращение ДГС (***DH***)).

Двойные гетероструктуры состоят из двух пассивных барьерных слоев и одного активного слоя. На рис.3.3 показана зонная диаграмма двойной гетероструктуры. Ширина запрещенной зоны активного слоя всегда меньше ширины запрещенной зоны барьерных слоев. В результате этого пассивные области являются *прозрачными* для излучения, исходящего из активной области. Поскольку пассивные слои, как правило, имеют сравнительно небольшую толщину, их практически всегда можно считать абсолютно прозрачными.

Перепоглощением света в активной области в месте инжекции носителей тока, расположенной под верхним контактом, можно также пренебречь. Поскольку ток носителей в активную область, как правило, имеет большую плотность, квазиуровни Ферми для электронов и дырок поднимаются к краям соответствующих зон, что и показано на рис.3.3. Поэтому при больших значениях инжекционного тока активная область является прозрачной для излучения с энергией, близкой к ширине запрещенной зоны.

Однако следует отметить, что равновесное состояние в активной области достигается в местах, достаточно удаленных от места инжекции тока. Именно в этих местах возможно поглощение излучения, генерируемого здесь же в активной области. Для компенсации оптических потерь из-за перепоглощения излучения внутри активной области, эти места должны обладать как можно более высоким внутренним квантовым выходом излучения.

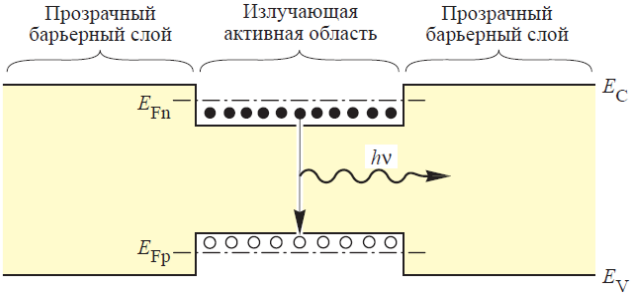

Рис.3.3. Распределение носителей тока в гетеропереходе .

На рис. 3.3 показано влияние гетеропереходов на распределение носителей. В двойных гетероструктурах барьерные слои ограничивают инжектированные носители в активной области. Поэтому величина области рекомбинации *определяется не диффузионной длиной, а толщиной активного слоя.*

Значения диффузионных длин обычно лежат в диапазоне 1…20 мкм, а размеры активной области в двойных гетероструктурах составляют 0.01…1.0 мкм. Это означает, что концентрация носителей в активной области двойных гетероструктур намного превышает концентрацию носителей в гомогенных переходах, где они распределены в интервале нескольких диффузионных длин. Из уравнения для скорости бимолекулярной излучательной рекомбинации:

$$R = B \cdot n \cdot p \qquad (3.7)$$

следует, что высокая концентрация носителей в активной области *увеличивает* скорость излучательной рекомбинации и снижает рекомбинационное время жизни. Поэтому все высокоэффективные светодиоды строятся на основе двойных гетероструктур, в частности на основе структур с квантовыми ямами.

Квантовая яма – это гетероструктура с размером области узкозонного полупроводника достаточно малой для проявления квантоворазмерных эффектов. Типичный размер квантовой ямы менее 100 Å.

Излучательная и безызлучательная рекомбинация

Рекомбинация электронов и дырок в полупроводниках бывает излучательной, т.е. с испусканием фотонов, и безызлучательной. В излучающих устройствах преобладающим является первый тип рекомбинации. Однако на практике безызлучательную рекомбинацию

никогда не удается довести до нуля. Таким образом, в полупроводниках всегда происходит конкуренция между двумя видами рекомбинации.

Внутренний квантовый выход излучения светодиода (или его ***внутренняяэффективность),*** определяется отношением числа рожденных в его активной области фотонов к числу инжектированных в нее электронов в единицу времени (секунду), т.е.:

$$\eta_{\text{int}} = \frac{P_{\text{int}} /(h\nu)}{I / e}, \tag{3.8}$$

где P_{int} – мощность оптического излучения из активной области светодиода, а I – ток инжекции. Внутренний квантовый выход идеального полупроводникового светодиода был бы равен 1, но квантовый выход реальных светодиодов всегда меньше 1.

Оценим внутренний квантовый выход полупроводников с центрами безызлучательной рекомбинации. Обозначим через τ_r – излучательное время жизни носителей, а через τ_{nr} – время жизни носителей в ходе безызлучательной рекомбинации. Тогда полная вероятность рекомбинации двух типов определяется суммой этих вероятностей:

$$\tau^{-1} = \tau_r^{-1} + \tau_{nr}^{-1}. \tag{3.9}$$

Относительную вероятность излучательной рекомбинации можно найти как отношение вероятности излучательной рекомбинации к суммарной вероятности. Таким образом, вероятность излучательной рекомбинации или ***внутренний квантовый выход излучения*** определяется следующим выражением:

$$\eta_{\text{int}} = \frac{\tau_r^{-1}}{\tau_r^{-1} + \tau_{nr}^{-1}}. \tag{3.10}$$

Внутренний квантовый выход равен отношению числа фотонов, испускаемых внутри полупроводникового материала, к числу электронно-дырочных пар, участвующих в актах рекомбинации. Отметим, что из-за проблем, связанных с поглощением света, далеко не все испущенные фотоны покидают пределы полупроводника.

Внешний квантовый выход излучения, коэффициент полезного действия (к.п.д.)

Желательно, чтобы все фотоны, испускаемые активной областью, выходили за пределы светодиода. Именно так и должно происходить в идеальных светодиодах, *внешний квантовый выход излучения* которых равен единице. Однако в реальных светодиодах часть фотонов все же остается внутри полупроводника: они могут быть

поглощены подложкой светодиода, либо металлической поверхностью контакта. Кроме того, не все фотоны могут покинуть полупроводниковую структуру из-за полного внутреннего отражения. Поэтому вводится такое понятие как ***коэффициент оптического вывода излучения***, определяемый как отношение числа фотонов, излученных светодиодом, к числу фотонов, образованных в активной области в единицу времени (секунду), т.е.:

$$\eta_{extraction} = \frac{P/(h\nu)}{P_{\text{int}}/(h\nu)}, \tag{3.11}$$

где P – мощность оптического излучения, выходящего за пределы светодиода. Этот параметр отражает качество светодиода. Без применения сложных и дорогих технологических процессов практически невозможно сделать светодиод с $\eta_{extraction} > 50\%$.

Внешний квантовый выход излучения светодиода определяется отношением числа фотонов, испущенных светодиодом к числу инжектированных электронов в единицу времени (секунду), т.е.:

$$\eta_{ext} = \frac{P/(h\nu)}{I/e} = \eta_{\text{int}} \cdot \eta_{extraction}. \tag{3.12}$$

Коэффициент полезного действия (к.п.д.) светодиода определяется как:

$$\eta_{power} = \frac{P}{I \cdot V}, \tag{3.13}$$

где $I \cdot V$ - электрическая мощность, подведенная к светодиоду.

<u>История создания полупроводниковых светодиодов</u>

В 1891 году Юджин Ачесон отработал процесс промышленного получения нового материала – карбида кремния (SiC), названного карборундом.

В 1907 году Генри Джозеф Раунд (1881-1966) при работе с кристаллами карборунда заметил испускаемое ими свечение.

В 1928 году Лосев опубликовал результаты своих исследований явления люминесценции, наблюдаемого в выпрямляющих диодах на основе SiC (переход металл – полупроводник). Он установил, что излучение света в одних диодах возникает только при их смещении в обратном направлении, а в других – при смещении как в прямом, так и обратном направлениях. Лосев установил, что излучение света никак не связано с сильным разогревом поверхности. Также он предположил, что явление люминесценции «очень похоже на процесс испускания холодных электронов». Лосев обнаружил, что процесс

появления и исчезновения свечения в SiC диодах происходил очень быстро, что делало возможным изготовление на их основе так называемых «световых реле».

К концу 60-х годов XX века были разработаны технологии получения пленок SiC и изготовления на их основе полупроводниковых устройств с p-n переходом. Диоды из карбида кремния являлись прародителями современных светодиодов голубого свечения, эффективность преобразования электрической энергии в оптическое излучение которых составляла всего 0.005%. В последующие десятилетия не удалось значительно улучшить характеристики светодиодов голубого свечения, что объясняется тем, что SiC относится к непрямозонным полупроводникам, которые отличаются очень низкой вероятностью межзонных оптических переходов. Лучшие SiC светодиоды излучали свет с длиной волны 470 нм и имели к.п.д. порядка 0.03%.

В 1954 году, после того как научились получать из расплавов монокристаллы GaAs, начался бум исследований полупроводниковых соединений типа $A^{III}B^{V}$. В 1962 году появилось сразу несколько публикаций о создании инфракрасных светодиодов (870…980 нм) и GaAs лазеров (Pankove, Berkeyheiser, 1962; Pankove, Massoulie, 1962; Hall и др., 1962; Nathan и др., 1962; Quist и др., 1962).

В начале 60-х годов научный коллектив, в состав которого входили известные ученые из IBM Thomas J. Watson Research Center, расположенного в Йорктаун Хейтс в часе езды к северу от Нью-Йорка: Джерри Вудалл, Ганс Руппрехт, Манфред Пилкухн, Маршалл Натан и др., провел большую исследовательскую работу по созданию GaAs и AlGaAs светодиодов и изучению их характеристик.

В 1962 году опубликовано сообщение о когерентном излучении видимого света, наблюдаемом на p-n переходе GaAsP при низкой температуре. В дальнейшем оказалось, что GaAsP светодиоды работают и при комнатной температуре.

В 1968 году кампания Monsanto Corporation построила завод, на котором стали изготавливать сравнительно недорогие GaAsP светодиоды. Этот год можно назвать началом эры твердотельных излучателей. Светодиодные кристаллы, выпускавшиеся Monsanto Corporation, представляли собой GaAsP p-n структуры, выращенные на GaAs подложках, излучающие фотоны с длиной волны, соответствующей красному диапазону видимого спектра.

Первые GaP светодиоды красного и зеленого свечения были созданы группой ученых под руководством Ральфа Логана в Bell Laboratories в Мюррее Хилл (Нью Джерси) в начале 1960-х годов. GaP

относится к непрямозонным полупроводникам, в которых вероятность межзонных переходов, происходящих с сохранением импульса, пренебрежимо мала, поэтому излучательная рекомбинация в них проходит, как правило, через примесные центры. Введение в GaP ***оптически активной изоэлектронной примеси***, например, N, позволяет значительно повысить вероятность излучательной рекомбинации в полупроводнике за счет того, что эта примесь создает в запрещенной зоне промежуточный энергетический уровень, с которого электрону гораздо легче рекомбинировать с дыркой.

В конце 1960-х годов была разработана технология получения GaP пластин из расплавов при высоких температурах и давлениях. Из таких пластин при помощи резки формировались подложки, точно такие же какие используются в настоящее время. При легировании GaP изоэлектронными примесями, содержащими N, такими как GaN, были изготовлены светодиоды зеленого свечения, к.п.д. которых превысил 0.6%.

Система материалов на основе AlInGaP подходит для получения яркого свечения в красном (626 нм), оранжевом (610 нм) и желтом (590 нм) спектральных диапазонах и в настоящее время является основной системой для изготовления светодиодов повышенной яркости, излучающих свет в данном интервале длин волн. Такая система материалов была разработана в Японии для лазеров, работающих в видимом диапазоне оптического спектра. Поскольку ширина запрещенной зоны InGaP составляет около 1.9 эВ (650 нм), этот материал может использоваться для изготовления лазеров, излучающих свет в красной области видимого спектра. Такие лазеры применяются, например, в лазерных указках и DVD проигрывателях.

Добавление Al к активной области InGaP позволяет сместить излучение в сторону более коротких длин волн, захватывая оранжевый и желтый спектральные диапазоны. Однако $(Al_xGa_{1-x})_{0.5}In_{0.5}P$ при $x \approx 0.53$ становится непрямозонным полупроводником, что приводит к сильному снижению его к.п.д на длинах волн, меньших или равных 600 нм. Следовательно, этот материал не подходит для изготовления высокоэффективных светодиодов, излучающих свет с длинами волн ниже 570 нм.

Первые AlInGaP лазеры появились в начале 1980-х годов, а развитие AlInGaP светодиодов началось в конце 1980-х. Дальнейшие усовершенствования AlInGaP светодиодов были связаны с созданием в активной области, состоящей из нескольких квантовых ям, распределенных отражателей Брэгга и технологии изготовления прозрачных GaP подложек (Kish, Fletcher, 1997).

Для создания светодиодов более коротковолнового излучения, в синей и зеленой области, требуется материал с более широкой запрещенной зоной. Таким материалом стал GaN. Летом 1971 года группой Панкова было опубликовано сообщение о первом явлении электролюминесценции, наблюдаемом на образце из пленки GaN. Исследуемый образец, состоявший из сильно легированного цинком GaN слоя с двумя поверхностными электродами, излучал свет голубого цвета с длиной волны 475 нм. После этого Панков с коллегами создали структуру из нелегированного GaN слоя (слоя n-типа), слоя сильно легированного Zn (диэлектрического слоя) и поверхностного контакта из In. Такой диод со структурой ***металл - диэлектрик – полупроводник*** (с МДП-структурой) был первым светодиодом на основе GaN, излучающим свет зеленого и голубого цвета.

В 1972 году удалось получить легированные Mg пленки GaN, излучающие на длине волны 430 нм. Такие светодиоды с МДП-структурой используются и в настоящее время. Следует отметить, что GaN пленки, даже легированные Mg, не обладают проводимостью p-типа, поэтому люминесценция в них протекает либо за счет инжекции неосновных носителей, либо за счет ударной ионизации диэлектрических слоев структур в сильном электрическом поле.

Работы в области исследований GaN были продолжены в группе Исаму Акасаки из Нагойи (Япония), которые в 1989 году продемонстрировали первый GaN светодиод со слоем, обладающим реальной проводимостью p-типа. Стойкие акцепторы Mg активировались при помощи ***облучения электронно-лучевым пучком***.

Первыми активировать стойкие акцепторы (Zn) в GaN удалось сотрудникам физического факультета МГУ им. М.В.Ломоносова М.В.Чукичеву и Г.В.Сапарину. В 1982 году при исследованиях катодолюминесценции пленок GaN, легированных Zn, они получили свечение образца GaN после облучения электронно-лучевым пучком. В начале 1990-х годов сотрудником японской фирмы NCIC Шуджи Накамурой было показано, что активировать примеси Mg в GaN слоях можно также в процессе ***высокотемпературного отжига*** готовых пленок. В настоящее время на основе GaN пленок, легированных Mg, изготавливаются все светодиоды и лазеры, содержащие нитриды. В 1992 году создан первый GaN светодиод с гомогенным p-n переходом, его к.п.д. порядка 1%. Позднее в этой же группе продемонстрированы InGaN светодиоды голубого и зеленого свечения с двойными гетероструктурами, к.п.д которых достиг уровня 10%.

Лекция 4. Фотометрия и цветные экраны

Возможности получения цветного изображения
Основы фотометрии.

Возможности получения цветного изображения

Появление в середине 1990 годов ярких светодиодов синего и зеленого свечения, а чуть раньше, в конце 1980-х годов, ярких светодиодов красного и желтого свечения, позволило покрыть весь видимый диапазон длин волн полупроводниковыми источниками излучения. В связи с этим стало возможно получать цветные изображения комбинацией различных светодиодов. Такие возможности сразу же нашли применения в индикации информации, например, на автомобильных дорогах – светофоры, информационные указатели и дорожные знаки, на железных дорогах – светофоры и информационные табло. Безусловно, прогресс в развитии полупроводниковых источников излучения – светодиодов, расширил диапазон применения оптических средств отображения и передачи информации.

Светодиод, состоящий из трех кристаллов – красного, зеленого и синего, позволяет получить практически любой цвет видимого диапазона оптического спектра. Подавая на кристаллы разные значения напряжения, т.е. пропуская через них разные токи, можно изменять не только цвета, а еще и оттенки того или иного цвета. Такой светодиод называется полноцветным.

Полноцветный светодиод в простейшем случае изготавливается следующим образом. Три кристалла – красный, зеленый и синий монтируются на общее основание, имеющее четыре вывода. Один вывод, например «+», является общим для трех кристаллов, а три контакта, в нашем случае «-», соединяются с контактами р-области трех кристаллов, по одному на каждый кристалл.

Пропуская различные токи через кристаллы легко менять цвет свечения такого светодиода. Также можно получить с помощью такого светодиода и белый цвет. Если взять приблизительное соотношение цветов в ТВ балансе белого цвета, которое составляет (исходя из относительной интенсивности белого цвета 1): красного - 0,25, зеленого - 0,65, синего – 0,1. Подбирая соответствующие значения токов через кристаллы этих трех цветов, можно получить необходимое соотношение интенсивности их свечения, и тем самым получить белый цвет. Такие светодиоды используются в различных целях, в том числе и в полноцветных экранах.

Для создания небольших полноцветных экранов можно использовать полноцветные светодиоды, описанные выше. Для создания больших экранов чаще используют просто несколько светодиодов в одном элементе экрана, например один красный, два зеленых и один синий. Возможно также использование готовых светодиодных кластеров, состоящих из нескольких светодиодов, или светодиодных кристаллов трех необходимых цветов. Такой кластер представляет собой единичный элемент экрана – пиксель. Тогда построение экрана будет заключаться в сборке таких пикселей, что существенно упрощает работу.

Основы фотометрии

Глаза человека являются основными «потребителями» информации, поступающей со светодиодов видимого спектра излучения. Внутренняя поверхность глазного яблока покрыта оболочкой, называемой сетчаткой, которая является светочувствительной частью глаза. Сетчатка состоит из стержневых и конических рецепторов (палочек и колбочек), а также нервов, передающих информацию в мозг. Стержневых рецепторов намного больше, чем конических, и они обладают более высокой светочувствительностью по сравнению с ними. Стержневые рецепторы воспринимают излучение всего видимого спектра, а конические – только определенной его части. Конические рецепторы можно разделить на три группы: рецепторы, воспринимающие излучение в красном, зеленом и голубом диапазонах видимого спектра, поэтому они так и называются рецепторами красного, зеленого и голубого цвета.

Разделяют три различных режима зрения: фотопической, скотопическое и мезопическое зрение. Для каждого из режимов работают разные группы рецепторов. Фотопическое зрение – это зрительное восприятие объектов человеческим глазом при высоких уровнях внешней освещенности (например, при дневном свете), когда в основном работают конические рецепторы. Скотопическое («сумеречное») зрение - это зрительное восприятие объектов человеческим глазом при низких уровнях внешней освещенности (например, ночью), когда в основном работают стержневые рецепторы, чувствительность которых намного выше, чем у конических рецепторов. Однако в этом режиме у человеческого глаза существенно снижается способность восприятия цвета. Ночью для человека все объекты теряют свои цвета и воспринимаются как объекты разных оттенков серого. Существует еще один режим зрения,

занимающий промежуточное положение между фотопическим и скотопическим зрением, называемый мезопическим зрением. Этот режим соответствует средней яркости, между диапазонами, соответственно, фотопического и скотопического зрения.

На рис. 4.1 приведены приблизительные функции спектральной чувствительности стержневых и трех типов конических рецепторов. Из рисунка видно, что ночное зрение (скотопический режим) по сравнению с дневным зрением (фотопический режим), слабее в красной части спектра и, соответственно, сильнее в голубом спектральном диапазоне. Следующие разделы будут посвящены, в основном, обсуждению режима фотопического зрения.

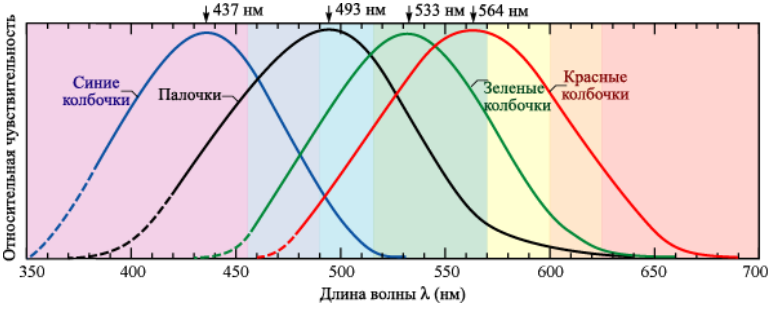

Рис. 4.1. Нормализованная спектральная чувствительность стержневых и конических рецепторов сетчатки человеческого глаза.

Радиометрические единицы характеризуют физические свойства электромагнитных излучений. Например, они используются для описания излучений в терминах следующих физических величин: количества фотонов, энергии фотонов, оптической мощности (часто называемой световым потоком). Однако когда речь идет о восприятии излучения человеческим глазом, использовать радиометрические единицы нельзя. Например, инфракрасное излучение не воспринимается глазом человека. Поэтому для характеристики свето- и цвето- восприятия глаза применяются не радиометрические, а фотометрические единицы.

Сила света является фотометрической величиной, характеризующей интенсивность излучения источника света по ее восприятию человеческим глазом. Сила света измеряется в канделах (кд), которые относятся к основным единицам измерения Международной системы единиц СИ. В настоящее время

используется следующее определение силы света: сила света монохроматического источника с мощностью излучения (1/683) Вт на длине волны 555 нм в пределах телесного угла1 стерадиан (ср), равна 1 канделе.

Единица кандела имеет большое историческое значение. Она произошла от более ранней единицы называемой свеча. Одна свеча равнялась силе света, излучаемого реальной свечкой, имеющей определенную конструкцию и размеры, а сила света от одной стандартной свечи равнялась 1.0 кд.

Фотометрической величиной, характеризующей мощность излучения, является световой поток. Световой поток определяется как сила света в полном телесном угле, равном 4π ср. Единицей измерения светового потока является люмен (лм). Следовательно можно сказать, что сила света равна светового потоку в пределах угла в 1 ср, или 1 кд = 1 лм/1 ср.

Восприятие цвета является субъективной величиной, которую нельзя измерить объективно. По этим причинам, Международная комиссия по освещению (МКО) стандартизировала цветовые измерения, введя понятия функций выравнивания цвета и цветового графика (цветовой диаграммы) (CIE, 1931).

Попробуем разобраться, как получают функции выравнивания цвета. Рассмотрим два источника света, расположенные рядом друг с другом. Один из источников монохромный, а цвет второго источника определяется составляющими света трех основных цветов: красного, зеленого и голубого (рис.4.2). При определенном подборе интенсивностей составляющих трех основных цветов (т.е. при их «выравнивании») человек будет воспринимать излучения двух источников одинаковыми. Три функции выравнивания цвета получаются из последовательности процедур выравнивания излучений от двух источников, в которых исследователь подбирает комбинации интенсивностей составляющих трех основных цветов одного из излучателей, соответствующих излучениям монохромных источников, работающих на разных длинах волн видимого спектра.

После этого полученный набор функций выравнивания цвета математически преобразуется в новый набор функций выравнивания цвета, в котором функция выравнивания зеленого цвета $\bar{y}(\lambda)$ выбирается так, чтобы ее числовые значения совпадали с функцией чувствительности человеческого глаза (функцией видности) $V(\lambda)$, т.е.

$$\bar{y}(\lambda)=V(\lambda). \qquad (4.1)$$

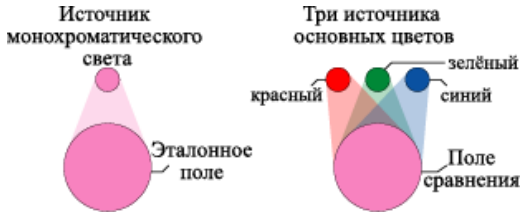

Рис. 4.2. Принцип выравнивания цветов: область, освещаемая тремя источниками основных цветов - красным, зеленым и голубым (обозначенная на рисунке как «поле сравнения»), выравнивается по цвету с областью, освещаемой эталонным источником монохромного света (обозначенная на рисунке как «эталонное поле»).

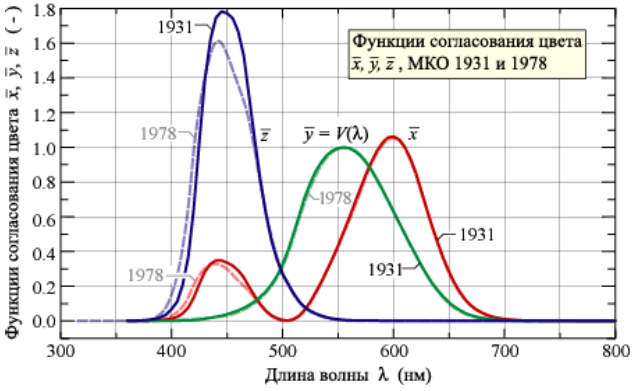

Рис. 4.3. $\bar{x}(\lambda), \bar{y}(\lambda), \bar{z}(\lambda)$ - функции выравнивания цвета, введенные МКО (1931) и МКО (1978). Функция $\bar{y}(\lambda)$ идентична функции чувствительности глаза $V(\lambda)$.

На рис.4.3 показаны функции выравнивания цвета $\bar{x}(\lambda), \bar{y}(\lambda), \bar{z}(\lambda)$, введенные МКО в 1931 и 1978 годах. Три функции выравнивания цвета $\bar{x}(\lambda), \bar{y}(\lambda), \bar{z}(\lambda)$ отражают факт, что человеческое зрение является трехцветным. Это означает, что цвет любого источника излучения может быть описан набором трех переменных $\bar{x}(\lambda), \bar{y}(\lambda), \bar{z}(\lambda)$, являющихся безразмерными величинами. Также следует отметить, что ни функции выравнивания цвета, ни цветовой график не являются

уникальными. Фактически существует несколько вариантов функций выравнивания цвета и цветовых графиков.

Отметим, что функции выравнивания цвета 1931года до сих пор являются официальным стандартом США.

Каждый цветовой оттенок, характеризующийся спектральной плотностью мощности излучения $P(\lambda)$, можно выразить набором трех параметров:

$$X = \int_{\lambda} \bar{x}(\lambda) \cdot P(\lambda) \mathrm{d}\lambda, \quad (4.2)$$

$$Y = \int_{\lambda} \bar{y}(\lambda) \cdot P(\lambda) \mathrm{d}\lambda, \quad (4.3)$$

$$Z = \int_{\lambda} \bar{z}(\lambda) \cdot P(\lambda) \mathrm{d}\lambda, \quad (4.4)$$

где X, Y и Z - параметры трёх основных цветов, соответствующие долям (например, мощности) каждой из составляющих трех основных (красного, зеленого и голубого) цветов, необходимых для получения излучения цвета $P(\lambda)$. Значения X, Y и Z указывают на количество красного, зеленого и голубого цвета в спектре $P(\lambda)$.

Поскольку между функциями чувствительности конических рецепторов сетчатки глаза и функциями выравнивания цветов наблюдается четко выраженное сходство (обе группы функций имеют три пиковых значения), можно считать, что каждый набор цветовых параметров X, Y и Z характеризует приблизительную (но не точную) степень возбуждения каждой группы конических рецепторов человеческого глаза, при попадании на них излучения от источника света со спектральной функцией $P(\lambda)$.

Из уравнений (4.2)…(4.4) видно, что цветовые параметры должны иметь размерность «Ватт». Но на практике эти параметры чаще всего задаются в виде безразмерных значений, для чего в соотношения вводятся дополнительные коэффициенты, например, сомножители впереди интегралов с размерностью «Ватт$^{-1}$». Однако если для получения безразмерных параметров цвета использовать соотношения, описанные ниже, дополнительные коэффициенты не нужны.

Для этого вводятся координаты цветности, определяемые в виде:

$$x = \frac{X}{X + Y + Z}, \quad (4.5)$$

$$y = \frac{Y}{X + Y + Z}. \quad (4.6)$$

Отсюда видно, что значение координаты цветности определяет степень возбуждения одной из групп цветовых рецепторов, нормализованную по величине суммарного возбуждения ($X+Y+Z$). Значение координаты z вычисляется при помощи аналогичного выражения:

$$z = \frac{Z}{X+Y+Z} = 1 - x - y. \qquad (4.7)$$

Отметим, что координату цветности z можно найти по известным значениям двух других координат, поэтому она является избыточной, и ее можно не использовать.

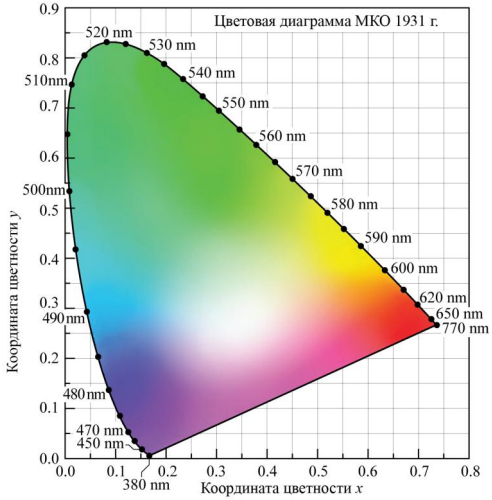

Рис. 4.4. Цветовой график МКО 1931 года. Монохроматические цвета расположены по периметру, а белый цвет находится в центре диаграммы.

На рис.4.4 показан цветовой график. Оттенки зеленого цвета соответствуют наибольшим значениям y, оттенки красного цвета – наибольшим значениям x, а оттенки голубого цвета – наибольшим значениям z, которые в соответствии с уравнением (17.7) расположены в области малых значений x и y или в области начала координат цветового графика.

На рис. 4.5 показано соответствие между цветами и их расположением на цветовом графике.

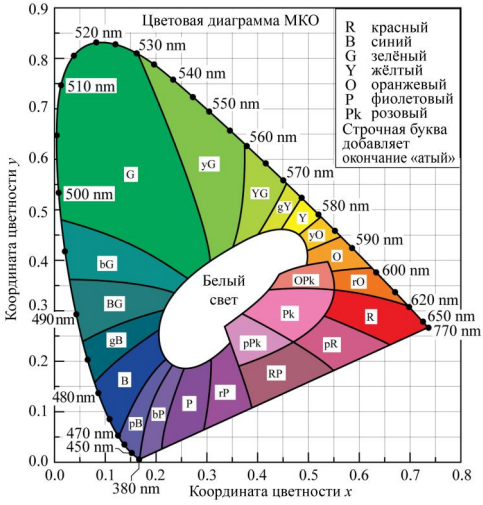

Рис. 4.5. Цветовой график МКО 1931 года с выделенными областями, соответствующими определенному цвету.

Монохроматические или чистые цвета расположены по периметру, а белый цвет находится в центре диаграммы цветности. Все цвета характеризуются координатами точек на цветовом графике.

Человек может различить цвета двух точек, если на цветовом графике между ними есть минимальное геометрическое расстояние. Это установил Мак Адам при анализе разности цвета в точках, расположенных на цветовом графике близко друг к другу. Цвета в пределах определенной (небольшой) области на цветовой диаграмме воспринимаются человеком как один цвет. Мак Адам показал, что эти области имеют форму эллипсов. Такие эллипсы, часто называемые эллипсами Мак Адама, показаны на рис.4.6. Из рисунка видно, что эллипсы в голубой и зеленой областях цветового графика довольно сильно различаются по размерам. Поэтому геометрическое расстояние между двумя точками на цветовой диаграмме не связано с разницей в цвете этих двух точек линейной зависимостью.

Общее количество различимых цветовых оттенков может быть найдено делением площади цветового графика на среднюю площадь эллипсов Мак Адама. Проведя такие вычисления, установили, что человек может различать 50000 различных цветовых оттенков. Если

учитывать все возможные изменения яркости, количество различаемых оттенков станет больше 10^6.

Монохромные источники ($\Delta\lambda \to 0$) располагаются по периметру цветового графика. А при увеличении ширины спектральной линии источника света, положение его цвета на цветовом графике смещается в сторону центра. Если ширина спектра излучения становится сравнима с шириной всего видимого диапазона спектра, источник считается источником белого света, который располагается вблизи центра цветовой диаграммы.

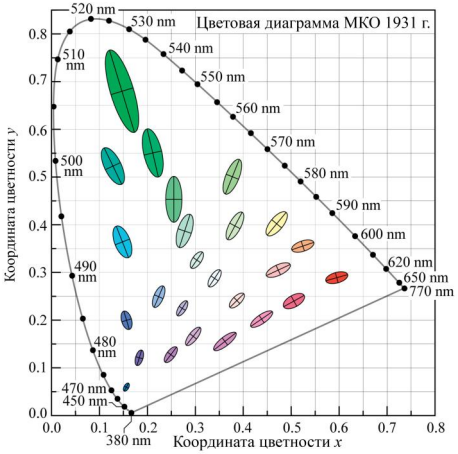

Рис. 4.6. Эллипсы Мак Адама, нанесенные на цветовой график МКО 1931 года.

Доминирующей (доминантной) длиной волны тестируемого источника света считается длина волны монохроматического цвета, расположенного на периметре цветового графика на самом коротком расстоянии от цвета излучения источника. Для определения доминирующей длины волны проводят прямую линию к периметру графика, проходящую через точку эталонного источника белого света и точку с координатами цветности (x, y), соответствующими тестируемому источнику света. Точка пересечения этой прямой с периметром цветового графика и будет определять доминирующую длину волны источника. На рис.4.7 схематично показана процедура нахождения доминирующей длины волны, а также типовые расположения светодиодов голубого, зеленого и красного свечений.

Чистота цвета или цветовая насыщенность источника света равна расстоянию на цветовом графике между точкой с координатами цветности (x, y) и точкой с координатами эталонного источника белого света, деленному на расстояние между точкой доминирующего цвета и точкой эталонного источника белого света. В качестве эталонного источника белого света часто используется равномерно-светящееся пятно (часто называемое пятном одинаковой энергии излучения). Для определения чистоты цвета используется выражение:

$$\text{чистота цвета} = \frac{a}{a+b} = \frac{\sqrt{(x-x_{ee})^2+(y-y_{ee})^2}}{\sqrt{(x_d-x_{ee})^2+(y_d-y_{ee})^2}}, \qquad (17.14)$$

где a и b показаны на рис.4.12, а (x, y), (x_{ee}, y_{ee}), (x_d, y_d) соответствуют координатам цветности тестируемого источника света, эталонного источника белого света и доминирующего цвета. Отсюда видно, что чистота цвета показывает, насколько удален тестовый источник излучения от центра цветового графика. Для монохромных источников ($\Delta\lambda \to 0$), расположенных по периметру цветового графика, чистота цвета, как правило, равна 100%, а для источников белого света она равна 0%.

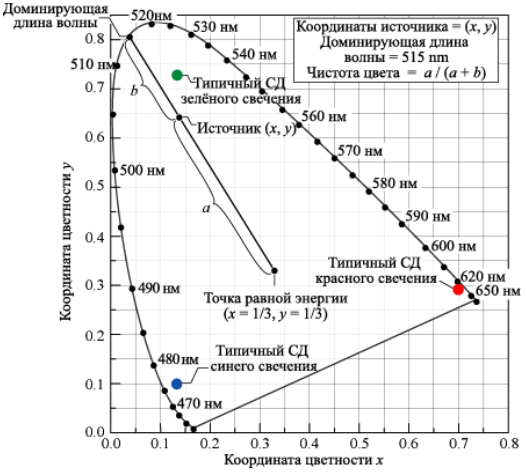

Рис. 4.7. Процедура нахождения доминирующей длины волны и определения чистоты цвета источника излучения с координатами цветности (x, y). Эталонный источник белого цвета – равномерно-светящееся пятно с координатами (x = 1/3, y = 1/3).

Отметим, что доминирующая длина волны и чистота цвета являются альтернативными параметрами, характеризующими положение излучателя на цветовом графике. На рис.4.7 показано расположение типовых светодиодов голубого (InGaN), зеленого (InGaN) и красного (AlInGaP) свечений. Чистота цвета светодиодов красного свечения очень высокая, часто близка к 100%. Светодиоды зеленого свечения из-за ненулевой ширины спектра излучения и сильной кривизны графика в зеленой части цветовой диаграммы обладают более низкой чистотой цвета. Доминирующая длина волны и чистота цвета являются достаточно интуитивными понятиями (в отличие от численных значений координат цветности), и поэтому ими бывает довольно удобно оперировать на практике.

Монохромные источники излучений ($\Delta\lambda \to 0$) располагаются по периметру цветового графика. Для человеческого глаза излучение светодиодов кажется монохромным, однако с точки зрения физики это не так, поскольку ширина спектральной линии светодиода является конечной величиной, приблизительно равной 1.8 *кТ*. Поэтому светодиоды располагаются не на самом периметре цветового графика, а вблизи него. Если источник излучает свет не одной длины волны, а в диапазоне длин волн, его координаты цветности сдвигаются ближе к центру графика.

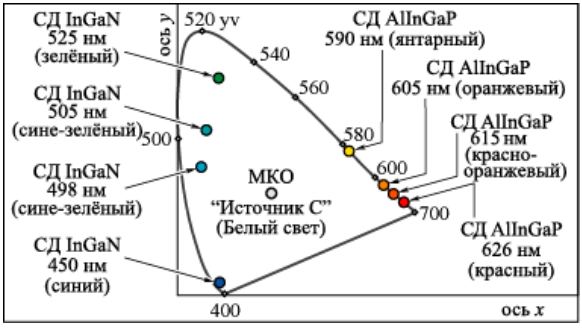

Рис. 4.8. Расположение на цветовом графике различных типов светодиодов.

Светодиоды красного и голубого свечения находятся на периметре цветового графика (рис. 4.8), а светодиоды сине-зеленного и зеленого свечения смещены от периметра ближе к центру диаграммы из-за конечной ширины их спектров и сильной кривизной графика в зеленой части цветовой диаграммы.

Лекция 5. Электромагнитная теория света

Уравнения Максвелла. Материальные уравнения. Гармонические световые волны в однородной изотропной линейной среде. Классическая осцилляторная модель среды.

Уравнения Максвелла

Распространение в пространстве световых волн, как и электромагнитных волн (полей) других диапазонов описывается уравнениями Максвелла

$$\begin{aligned} &rot\vec{H} - \frac{\partial \vec{D}}{\partial t} = \vec{J}; \\ &rot\vec{E} + \frac{\partial \vec{B}}{\partial t} = 0; \\ &div\vec{D} = \rho; \\ &div\vec{B} = 0; \end{aligned} \qquad (5.1)$$

Здесь $\vec{E}$, $\vec{H}$ и $\vec{D}$, $\vec{B}$ – напряженности и индукции соответственно электрического и магнитного полей, $\vec{J}$ и ρ – плотности тока проводимости и зарядов.

Первое из этих уравнений может быть преобразовано к виду, аналогичному уравнению непрерывности в гидродинамике. Для этого возьмем *div* от обеих частей уравнения. Учитывая, что *div rot* = 0, а также возможность изменения порядка дифференцирования по времени и координатам, имеем $div\vec{J} = -\frac{\partial}{\partial t} div\vec{D}$, и, воспользовавшись третьим уравнением Максвелла из (5.1), получим

$$\frac{\partial \rho}{\partial t} + div\vec{J} = 0. \qquad (5.2)$$

Из уравнения (5.2) следует, что изменение заряда, а точнее его плотности, в окрестности любой точки может происходить только при появлении дополнительных токов.

Если все величины, связанные с полем, не зависят от времени и отсутствуют токи ($\vec{J} = 0$), то такое поле называется *статическим*, если же все величины не зависят от времени, но присутствуют токи, то такое поле называется *стационарным*.

Система уравнений (5.1) не содержит характеристик среды и в этом смысле является универсальной, применимой для любых сред – однородных и неоднородных, изотропных и анизотропных, стационарных и нестационарных, в отсутствии или при наличии нелинейных эффектов. Однако эта система не замкнута и дополняется

материальными уравнениями, которые устанавливают связь между $\vec{E}$, $\vec{H}$, $\vec{D}$, $\vec{B}$, $\vec{J}$, ρ. В материальных уравнениях и находят отражение особенности той или иной конкретной среды.

Будем рассматривать в дальнейшем распространение света в диэлектрических средах, в которых $\vec{J}=0$, $\rho=0$, что означает отсутствии свободных зарядов в среде.

Векторы электрической и магнитной индукции $\vec{D}(\vec{r},t)$ и $\vec{B}(\vec{r},t)$ возникают как отклик среды на электрическое и магнитное поля $\vec{E}(\vec{r},t)$ и $\vec{H}(\vec{r},t)$, распространяющиеся в среде, и связаны с ними следующими материальными уравнениями

$$\vec{D}=\varepsilon_0\vec{E}+\vec{P} \tag{5.3}$$

$$\vec{B}=\mu_0\vec{H}+\vec{M} \tag{5.4}$$

где ε_0 и μ_0 – диэлектрическая и магнитная постоянные вакуума, $\vec{P}(\vec{r},t)$ и $\vec{M}(\vec{r},t)$ – индуцированные электрическая и магнитная поляризации. В оптике в большинстве случаев световые волны распространяются в немагнитных веществах $\vec{M}(\vec{r},t)=0$.

Уравнения Максвелла могут быть использованы для получения уравнения, описывающего распространение света в веществе. Применив операцию ротора к первому уравнению системы (5.1), используя второе уравнение этой системы и уравнения (5.2) и (5.3), можно исключить переменные $\vec{B}(\vec{r},t)$ и $\vec{D}(\vec{r},t)$, оставив только переменные $\vec{E}(\vec{r},t)$ и $\vec{P}(\vec{r},t)$:

$$\vec{\nabla}\times\vec{\nabla}\times\vec{E}=-\frac{1}{c^2}\frac{\partial^2\vec{E}}{\partial t^2}-\mu_0\frac{\partial^2\vec{P}}{\partial t^2} \tag{5.5}$$

где используется соотношение $\mu_0\varepsilon_0= 1/c^2$, c – скорость света в вакууме.

Для математического описания взаимодействия света с веществом нужно ввести связь между индуцированной поляризацией $\vec{P}(\vec{r},t)$ и электрическим полем $\vec{E}(\vec{r},t)$. Вообще говоря, чтобы определить $\vec{P}(\vec{r},t)$, нужно использовать квантово-механическую теорию. Однако такой подход часто бывает необходим только тогда, когда частота оптического поля близка к резонансным частотам среды. Вдали от резонансных частот, для связи $\vec{P}(\vec{r},t)$ и $\vec{E}(\vec{r},t)$ можно использовать самые общие феноменологические соотношения.

По характеру функциональной зависимости $\vec{P}(\vec{r},t)$ от $\vec{E}(\vec{r},t)$ можно провести классификацию материальных сред. Среды, в которых зависимость поляризации от поля является локальной и безынерционной, т.е. значение поляризации среды в некоторой точке в некоторый момент времени определяется значением поля в той же самой точке и в тот же момент, называются ***недиспергирующими***. Нелокальность отклика приводит ***пространственной*** дисперсии среды, а его инерционность, т.е. запаздывание поляризации относительно поля – к ***временной*** или ***частотной*** дисперсии. ***Линейными*** называются среды, в которых зависимость поляризации от поля выражается линейным оператором, в частности, линейным запаздывающим функционалом или тензорным оператором. Соответственно, ***нелинейными*** называются среды, для которых эта зависимость нелинейна. ***Изотропными*** называют среды, в которых поляризация ориентирована параллельно электрическому полю. Наконец, в ***анизотропных*** средах вектор поляризации, вообще говоря, не параллелен вектору электрического поля.

Материальные уравнения

Связь между индуцированной поляризацией $\vec{P}(\vec{r},t)$ и электрическим полем $\vec{E}(\vec{r},t)$ определяют материальные уравнения. Будем исходить из того, что отклик любого диэлектрика на световое воздействие в общем случае нелинейный и не мгновенный, не локальностью отклика пренебрежем. С теоретической точки зрения возникновение нелинейного отклика связано с ангармоническим движением связанных электронов и ионов при воздействии приложенного поля. В результате индуцированная поляризация $\vec{P}(\vec{r},t)$ электрических диполей может быть выражена в следующем виде:

$$\vec{P}(\vec{r},t) = \vec{P}_L(\vec{r},t) + \vec{P}_{NL2}(\vec{r},t) + \vec{P}_{NL3}(\vec{r},t) + ... \qquad (5.6)$$

где $\vec{P}_L(\vec{r},t)$ линейная часть поляризации а $\vec{P}_{NLj}(\vec{r},t)$ – нелинейная часть поляризации j-го порядка. Главный вклад в $\vec{P}(\vec{r},t)$ вносит линейная восприимчивость, определяющая возникновение $\vec{P}_L(\vec{r},t)$. Она определяет показатель преломления n и постоянную затухания α. Линейная часть связана с электрическим полем следующим соотношением:

$$\vec{P}_L(\vec{r},t) = \varepsilon_0 \int_{-\infty}^{t} \hat{\chi}^{(1)}(t-t') \cdot \vec{E}(\vec{r},t')dt' \qquad (5.7)$$

где $\hat{\chi}^{(1)}(t-t')$ – в общем случае тензор размерности 3x3.

С нелинейностью второго порядка $\vec{P}_{NL2}(\vec{r},t)$ связаны такие эффекты, как генерация второй гармоники и генерация суммарной частоты. Однако эта восприимчивость не равна нулю только для сред, в которых на молекулярном уровне отсутствует симметрия инверсии. Так как в кварцевых стеклах молекула SiO_2 обладает центром симметрии в оптических волокнах, обычно не наблюдаются эффекты второго порядка. Тем не менее, слабые нелинейные эффекты второго порядка могут возникать из-за электрических квадрупольных и магнитных дипольных моментов. Примеси внутри сердцевины волокна могут также при определенных условиях приводить к генерации второй гармоники. Нелинейные эффекты выше третьего порядка обычно малы, и обычно ими можно пренебречь.

При анализе распространения световых сигналов в кварцевых волокнах, таким образом, достаточно рассмотреть вклад линейной поляризации и нелинейной поляризации третьего порядка (кубичной нелинейности). В этом случае нелинейная часть связана с электрическим полем следующим соотношением:

$$\vec{P}_{NL3}(\vec{r},t) = \varepsilon_0 \int_{-\infty}^{t} dt_1 \int_{-\infty}^{t} dt_2 \int_{-\infty}^{t} dt_3 \times \hat{\chi}^{(3)}(t-t_1, t-t_2, t-t_3) \vdots \vec{E}(\vec{r},t_1)\vec{E}(\vec{r},t_2)\vec{E}(\vec{r},t_3) \tag{5.8}$$

где $\hat{\chi}^{(3)}(t-t_1, t-t_2, t-t_3)$ – в общем случае тензор 3 ранга. Соотношения (5.7), (5.8) справедливы в дипольном приближении, когда предполагается, что отклик среды является локальным, но не мгновенными.

Ввиду сложности уравнений (5.6)-(5.8) обычно используют дополнительные упрощающие предположения. Наиболее общее упрощение состоит в том, что нелинейная поляризация $\vec{P}_{NLj}$ в (5.6) считается малым возмущением полной индуцированной поляризации. Такое приближение оправданно, так как в оптических системах связи нелинейные эффекты относительно слабы. Более того, часто можно ограничиться рассмотрением распространения световых сигналов в линейном приближении.

Гармонические световые волны в однородной изотропной линейной среде

В материальной однородной линейной среде также как и в вакууме существуют точные решения волнового уравнения в виде гармонических волн. (Отметим, что решения в виде гармонических волн в нелинейных средах, являются приближенными, хотя в некоторых случаях такие приближенные решения очень мало

отличается от точных). Решение в виде физически реальных волн может быть записано следующим образом:

$$\vec{E}_{RE}(\vec{r},t) = \vec{E}_{0RE}(r)\cos[\omega t + \varphi(r)], \tag{5.9}$$

где $\vec{E}_{0RE}(r)$ – амплитуда электрического поля, ω – ***угловая частота***, $\varphi(r)$ – ***фаза*** электрического поля. Амплитуда и фаза электрического поля зависят от пространственной координаты r. Угловая частота связана с ***частотой*** ν и ***периодом*** T выражением:

$$\omega = 2\pi\nu = 2\pi / T \tag{5.10}$$

Для описания гармонических волн удобно использовать комплексную форму, которая существенно упрощает вычисления.

$$\vec{E}_{RE}(\vec{r},t) = \mathrm{Re}[\tilde{\vec{E}}(r,t)] = \mathrm{Re}[\tilde{\vec{E}}_Q(r)\exp(i\omega t)], \tag{5.11}$$

где $\tilde{\vec{E}}_Q(r)$ – ***комплексная амплитуда электрического поля***.

Комплексная амплитуда может быть выражена через действительную амплитуду и фазу гармонического колебания следующими двумя эквивалентными выражениями:

$$\tilde{\vec{E}}_Q(\vec{r}) = \vec{E}_{0RE}(r)\exp(i\varphi)], \tag{5.12}$$

$$\tilde{\vec{E}}_Q(\vec{r}) = \vec{E}_{0RE}(r)\cos(\varphi) + i\vec{E}_{0RE}(r)\sin(\varphi), \tag{5.13}$$

Подстановка выражения (5.11) для гармонической волны в уравнение (5.5) приводит к уравнению Гельмгольца для световых волн в диэлектрике:

$$\nabla^2\tilde{\vec{E}}_{0Q}(\vec{r},\omega) = \varepsilon(\omega)\frac{\omega^2}{c^2}\tilde{\vec{E}}_{0Q}(\vec{r},\omega) \tag{5.14}$$

где $\tilde{\vec{E}}_{0Q}(\vec{r},\omega)$ – комплексная амплитуда электрического поля гармонической световой волны. Относительная диэлектрическая проницаемость, зависящая от частоты, определяется следующим образом:

$$\varepsilon(\omega) = 1 + \chi^{(1)}_{\omega}(\omega), \tag{5.15}$$

где $\chi^{(1)}_{\omega}(\omega)$ --- Фурье-преобразование функции $\chi^{(1)}(t)$. Вообще $\chi^{(1)}$, а следовательно, и $\varepsilon(\omega)$ - комплексные величины. По определению показатель преломления $n(\omega)$ и коэффициент поглощения $\alpha(\omega)$ связаны с действительной и мнимой частями $\varepsilon(\omega)$:

$$\varepsilon(\omega) = (n + i\alpha c/2\omega)^2. \tag{5.16}$$

С помощью выражений (5.15) и (5.16) $n(\omega)$ и $\alpha(\omega)$ можно выразить через $\chi^{(1)}_{\omega}(\omega)$:

$$n(\omega) = 1 + (1/2)\,\mathrm{Re}[\chi^{(1)}_{\omega}(\omega)] \tag{5.17}$$

$$\alpha(\omega) = \frac{\omega}{nc}\,\mathrm{Im}[\chi_{\omega}^{(1)}(\omega)] \qquad (5.18)$$

здесь Re и Im обозначают соответственно действительную и мнимую части комплексного числа.

Простейшее, но очень важное решение уравнения (5.5) в однородной среде представляет собой плоскую монохроматическую волну, распространяющуюся в направлении оси z:

$$\vec{E}_{RE}(z,t) = \mathrm{Re}[\tilde{\tilde{E}}_{0Q}(0)\exp(i(\omega t - k_0 n(\omega) z) - (\alpha/2)z)], \qquad (5.19)$$

В отличие от плоской монохроматической волны в вакууме фазовая скорость распространения плоской монохроматической волны в материальной среде $v_P = c/n(\omega)$ зависит от частоты, а ее амплитуда уменьшается по мере распространения с коэффициентом затухания по амплитуде $\alpha/2$ (коэффициент затухания по интенсивности равен α). Волны, распространяющиеся в однородных средах, характеризуются определенной связью между длиной волны $\lambda = 2\pi/k$ и частотой волны ω. Введенный параметр k называется волновым числом. Его взаимосвязь с частотой волны ω

$$k(\omega) = \omega n(\omega)/c = \omega[\varepsilon(\omega)]^{1/2}/c, \qquad (5.20)$$

или обратная ей зависимость $\omega(k)$ называется ***законом дисперсии***, а удовлетворяющие ей волны называются ***свободными*** или ***нормальными***.

В прозрачных однородных средах можно пренебречь мнимой частью $\varepsilon(\omega)$ ($\mathrm{Im}(\varepsilon) = 0$). Тогда в дальнейшем $\varepsilon(\omega)$ можно заменить на $n^2(\omega)$, а потери можно будет включить в рассмотрение позднее, применяя метод возмущений. При таких упрощениях волновое уравнение принимает следующую форму:

$$\nabla^2 \tilde{\tilde{E}}_Q + n^2 \frac{\omega^2}{c^2}\tilde{\tilde{E}}_Q = 0, \qquad (5.21)$$

Для получения вида дисперсионных зависимостей $\omega(k)$ надо найти явный вид функции $\chi^{(1)}(t)$ или ее Фурье-образ $\chi_{\omega}^{(1)}(\omega)$. Эти две функции связаны прямым и обратным преобразованием Фурье:

$$\hat{\chi}^{(1)}(\omega) = \int_{-\infty}^{\infty} \hat{\chi}^{(1)}(t') \cdot \exp(-i\omega t')dt' \qquad (5.22)$$

$$\hat{\chi}^{(1)}(t') = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{\chi}^{(1)}(\omega) \cdot \exp(i\omega t')d\omega \qquad (5.23)$$

Так как отклик среды не может опережать воздействие на нее, то $\chi^{(1)}(t<0)=0$, а выражение (5.22) принимает вид:

$$\hat{\chi}^{(1)}(\omega) = \int_0^{\infty} \hat{\chi}^{(1)}(t') \cdot \exp(-i\omega t')dt' \qquad (5.24)$$

Зависимость поляризации от поля (1.2.9) можно записать в виде

$$\vec{P}_L(\vec{r},t) = \varepsilon_0 \int_0^{\infty} \hat{\chi}^{(1)}(t') \cdot \vec{E}(\vec{r},t-t')dt' \qquad (5.25)$$

Формула (5.25) носит название ***интеграла Дюамеля***. Согласно этой формуле, отклик линейной системы есть линейный запаздывающий функционал относительно входного воздействия. Математически в этой формуле $\chi^{(1)}(t')$ – функция Грина, зависящая только от свойств среды и связанная с линейной оптической восприимчивостью среды $\chi^{(1)}(\omega)$ формулой (5.24).

Классическая осцилляторная модель среды

В приближении линейной изотропной однородной среды ее оптическую поляризацию $\vec{P}$, имеющую смысл дипольного момента единицы объема, можно представить в виде:

$$\vec{P} = N\vec{p} \qquad (5.26)$$

где N – число частиц (атомов или молекул) в единице объема, а $\vec{p}$ – дипольный момент отдельной частицы. Частицы среды могут быть либо нейтральными, либо обладать собственным дипольным моментом. Во втором случае в изотропном веществе ориентация диполей в отсутствии внешнего поля случайная и в целом поляризация среды отсутствует. Однако, если приложено внешнее поле, то происходит либо поляризация частиц, либо поворот уже имеющихся дипольных моментов частиц под действием электрического полю, и у единицы объема появляется средний дипольный момент.

В линейном приближении дипольный момент молекулы пропорционален электрическому полю

$$\vec{p} = \alpha_A \vec{E}_D, \qquad (5.27)$$

где $\vec{E}_D$ – действующее на молекулу электрическое поле. В разреженном газе, содержащем N частиц в единице объема, взаимодействием дипольных моментов частиц можно пренебречь ($\vec{E}_D = \vec{E}$) и объемная поляризация среды $\vec{P}$ определяется выражением

$$\vec{P} = N\vec{p} = N\alpha_A \vec{E} \qquad (5.28)$$

Для газов относительная диэлектрическая проницаемость с учетом (5.28) равна:

$$\varepsilon = 1 + \chi_{\omega}^{(1)}(\omega) = 1 + (\vec{P}/\varepsilon_0 \vec{E}) = 1 + N\alpha_A/\varepsilon_0 . \qquad (5.29)$$

В случае конденсированных веществ, в частности твердых диэлектриков, необходимо учитывать влияние, оказываемое на степень поляризации каждой отдельной частицы окружающими ее частицами. При использовании простейшего приближения, которое оказывается точным для идеальной кубической решетки, полагают, что каждая поляризуемая частица представляет собой сферическую замкнутую полость в однородном диэлектрике. При этом под действием среднего (внешнего) поля $\vec{E}$ локальное поле (действующее на частицу) увеличивается $E_D = \vec{E} + \vec{P}/3\varepsilon_0$. Физически это означает, что на каждую частицу конденсированного вещества действуют помимо внешнего поля $\vec{E}$ еще и поля, создаваемые дипольными моментами окружающих ее частиц. Таким образом, в конденсированном веществе в соответствии с (5.27) поляризация среды определяется следующим неявным выражением:

$$\vec{P} = N\vec{p} = N\alpha_A \vec{E}_D = N\alpha_A(\vec{E} + \vec{P}/3\varepsilon_0). \quad (5.30)$$

Из (3.30) несложно получить явное выражение для зависимости $\vec{P}(\vec{E})$:

$$\vec{P} = N\alpha_A \vec{E}/[1-(N\alpha_A/3\varepsilon_0)]. \quad (5.31)$$

Таким образом, в случае твердых диэлектриков выражение для относительной диэлектрической проницаемости будет иметь вид:

$$\varepsilon = 1 + \chi_\omega^{(1)}(\omega) = 1 + (\vec{P}/\varepsilon_0 \vec{E}) = 1 + \frac{N\alpha_A/\varepsilon_0}{1 - N\alpha_A/3\varepsilon_0}. \quad (5.32)$$

Этот результат иногда выражают в иной форме, предложенной Моссотти:

$$(\varepsilon - 1)/(\varepsilon + 2) = N\alpha_A/3\varepsilon_0. \quad (5.33)$$

Из (5.33) легко выразить в явном виде зависимость поляризуемости молекулы конденсированного вещества от диэлектрической проницаемости, которая получила название формулы Лорентц-Лоренца:

$$\alpha_A = \frac{3\varepsilon_0}{N}\frac{(\varepsilon - 1)}{(\varepsilon + 2)}. \quad (5.34)$$

Рассмотрим классическую «осцилляторную» модель среды, предложенную Лоренцем. Согласно этой модели вещество состоит из нейтральных атомов, дипольный момент $\vec{p}$ которых определяется смещением $\vec{x}$ электронов из состояния равновесия (относительно атомного ядра):

$$\vec{p} = e\vec{x} \quad (5.35)$$

где e – заряд электрона. При воздействии электрического поля электромагнитной волны уравнение движения электрона имеет вид:

$$\ddot{x} + \Gamma\dot{x} + \omega_{0K}^2 x = (e/m)E_D \exp(-i\omega t), \quad (5.36)$$

где m – масса электрона, ω_0 – собственная частота колебаний электрона, параметр Γ – описывает затухание колебаний. Решение уравнения (5.36) имеет вид:

$$\tilde{x} = (e/m)\frac{E_D}{\omega_0^2 - \omega^2 + i\omega\Gamma}, \tag{5.37}$$

Поскольку $\vec{P} = N\vec{p} = Nex$ для поляризации газа получаем выражение

$$\vec{P} = \frac{Ne^2}{m}\frac{E}{\omega_0^2 - \omega^2 + i\omega\Gamma} \tag{5.38}$$

Сравнив формулы, находим выражение для линейной оптической восприимчивости газа в модели Лоренца:

$$\tilde{\chi}_{\omega}^{(1)}(\omega) = \frac{Ne^2}{\varepsilon_0 m}\frac{1}{\omega_0^2 - \omega^2 + i\omega\Gamma} \tag{5.39}$$

$$\tilde{\varepsilon}(\omega) = 1 + \frac{Ne^2}{\varepsilon_0 m}\frac{1}{\omega_0^2 - \omega^2 + i\omega\Gamma} \tag{5.40}$$

Для разреженных газов нетрудно получить также функцию импульсного отклика подставив (5.39) в (5.23)

$$\chi^{(1)}(\tau) = \left.\begin{array}{ll} 0, & \tau \le 0, \\ \dfrac{Ne^2}{m\sqrt{\omega_0^2 - \Gamma^2/4}}\exp(-\dfrac{\Gamma}{2}\tau)\sin(\sqrt{\omega_0^2 - \Gamma^2/4}\tau), & \tau \ge 0, \end{array}\right\} \tag{5.41}$$

Аналогично для конденсированного вещества с учетом (5.33) получаем следующее неявное выражение для относительного показателя преломления

$$\frac{(\tilde{\varepsilon} - 1)}{(\tilde{\varepsilon} + 2)} = \frac{Ne^2}{3\varepsilon_0 m}\frac{1}{\omega_0^2 - \omega^2 + i\omega\Gamma}. \tag{5.42}$$

или следующее явное выражение, получаемое из (5.32)

$$\tilde{\varepsilon}(\omega) = 1 + \frac{Ne^2/m\varepsilon_0}{\omega_0^2 - \omega^2 + i\omega\Gamma - (Ne^2/3m\varepsilon_0)} \tag{5.43}$$

Из (5.43) и (5.40) видно, что в случае конденсированных диэлектриков сохраняется резонансная форма зависимостей $\tilde{\varepsilon}(\omega)$ и изменяется только резонансная частота ω_R

$$\omega_R = \sqrt{\omega_0^2 - (Ne^2/3m\varepsilon_0)}, \tag{5.44}$$

Если учесть все возможные резонансы и ввести коэффициенты g_k, называемые силой осциллятора, получим

$$\tilde{\varepsilon}(\omega) = 1 + (Ne^2/m\varepsilon_0)\sum_k \frac{g_k}{\omega_{Rk}^2 - \omega^2 + i\omega\Gamma} \tag{5.45}$$

Из формулы (5.45) следует связь между действительной и мнимой частями диэлектрической проницаемости.

Лекция 6. Распространение световых сигналов

Скорость распространения световых сигналов. Формула Селмейера для прозрачных диэлектриков. Распространение световых импульсов. Волновое уравнение для огибающей светового импульса. Векторные волны. Поляризация.

Скорость распространения световых сигналов (групповая скорость)

Амплитудно-модулированные световые волны, а также световые сигналы, в которых осуществляется модуляция других параметров световой волны, по определению распространяются с групповой скоростью, т.е. со скоростью распространения огибающей (сигнала). Однако в общем случае сигнал по мере распространения искажается и может возникнуть неоднозначность в определении скорости распространения сигнала. В приближении узкополосного сигнала, когда ширина спектра сигнала много меньше несущей частоты, групповая скорость определяется следующим выражением:

$$v_G = d\omega / dk \qquad (6.1)$$

В среде без дисперсии, т.е. в такой среде, где нет зависимости $n = n(\omega)$, фазовая и групповая скорости равны (известна только одна среда без дисперсии - вакуум). Однако, если в среде присутствует дисперсия, то эти скорости различны, а если среда является ещё и анизотропной, то векторы этих скоростей, в общем случае, неколлинеарные.

Следует отметить, что именно групповая скорость является скоростью передачи информации в оптических системах связи. В отличие от фазовой скорости групповая скорость световой волны не может быть больше скорости распространения света в вакууме (за исключением сильно поглощающих или усиливающих сред, которые мы рассматривать не будем).

Понятие групповой скорости можно пояснить на простом примере распространения бигармонической световой волны вида (см. рис.6.1.):

$$E = A\cos(\omega_1 t - k_1 z) + A\cos(\omega_2 t - k_2 z). \qquad (6.2)$$

Правую часть (6.1) можно представить в виде:

$$E = 2A\cos\left[\frac{\omega_1 - \omega_2}{2} t - \frac{k_1 - k_2}{2} z\right]\cos\left[\frac{\omega_1 + \omega_2}{2} t - \frac{k_1 + k_2}{2} z\right]. \qquad (6.3)$$

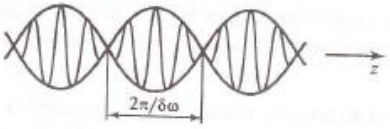

Рис. 6.1. Бигармоническая световая волна.

Формула (6.2) показывает, что распространение бигармонической волны характеризуется двумя скоростями: фазовой скоростью $v_P = (\omega_1 + \omega_2)/(k_1 + k_2)$ и групповой скоростью $v_G = (\omega_1 - \omega_2)/(k_1 - k_2)$. Фазовая скорость характеризует скорость быстро осциллирующей компоненты (второй косинус в (6.2)), т.е. несущей световой волны. Групповая скорость характеризует скорость распространения медленно осциллирующей компоненты (первый косинус в (6.2)), т.е. огибающей световой волны.

Фазовая скорость распространения электромагнитных волн в среде $v_P = c/\sqrt{\varepsilon\mu}$. Для прозрачных веществ $\varepsilon > 1, \mu = 1$, а фазовая скорость света в среде v_P меньше скорости света в вакууме c. Величина $n = \sqrt{\varepsilon\mu} \approx \sqrt{\varepsilon}$ называется фазовым показателем преломления среды. В некоторых типах волноводов фазовая скорость световых волн может быть больше скорости света.

Формула Селмейера для прозрачных диэлектриков

В интересной для оптической связи области – области прозрачных диэлектриков – затухание должно быть очень малым. Поэтому можно ограничиться рассмотрением частот, расположенных далеко от резонансных. В этой области мнимой частью показателя преломления можно пренебречь и с достаточно высокой степенью точности можно использовать следующее выражение для фазового показателя преломления:

$$n^2 - 1 = (Ne^2/m\varepsilon_0)\sum_k \frac{g_k}{\omega_{Rk}^2 - \omega^2} \qquad (6.4)$$

Изменение показателя преломления можно преобразовать в зависимость от длины волны

$$n^2 - 1 = \sum_k \frac{G_k \lambda^2}{\lambda^2 - \lambda_{Rk}^2} \qquad (6.5)$$

где $G_k = (Ne^2 \lambda_{Rk}^2 g_k)/(4\pi^2 mc^2 \varepsilon_0)$. Обычно изменение показателя преломления оптических материалов выражают именно в виде выражения (6.2), известного как дисперсионная формула Селмейера.

Очень хорошее соответствие теоретических и экспериментальных результатов наблюдается при учете трех членов дисперсионной формулы, два из которых соответствуют электронным резонансам в ультрафиолетовой области спектра, а один возникает вследствие молекулярного резонанса в инфракрасной области.

$$n^2 - 1 = \frac{G_1\lambda^2}{\lambda^2 - \lambda_{R1}^2} + \frac{G_2\lambda^2}{\lambda^2 - \lambda_{R2}^2} + \frac{G_3\lambda^2}{\lambda^2 - \lambda_{R3}^2} \qquad (6.6)$$

В 1965 г. сотрудник Национального Бюро стандартов США Малитсон (Malitson) определил с высокой точностью показатель преломления объемных образцов из чистого кварца в диапазоне длин волн от 0,2 до 4 мкм. Полученные результаты он представил в виде дисперсионного уравнения Селмейера (6.6) с тремя членами со следующими значениями коэффициентов и резонансных длин волн: G_1=0,6961663, G_2 = 0,4079426, G_3 = 0,8974794, λ_1 = 0,0684043 мкм, λ_2= 0.1162414 мкм и λ_3 = 9.896161 мкм.

Любая огибающая светового излучения или сигнал распространяются не с фазовой $v_P = c/n(\omega)$, а с групповой скоростью $v_G = d\omega/dk = c/n_G(\omega)$. Групповой показатель преломления связан с фазовым показателем преломления следующим выражением:

$$n_G = n - \lambda\frac{dn}{d\lambda} \qquad (6.7)$$

Зависимости фазового и группового показателей преломления от частоты для плавленого кварца приведены на рис.6.2.

В оптике слово «дисперсия» обычно связывают с зависимостью показателя преломления от частоты $n(\omega)$, которую можно охарактеризовать величиной $dn/d\lambda$. В оптических системах связи этот же термин связывают с явлением уширения световых импульсов после их прохождения через дисперсионную среду. Как будет показано ниже, численно величина уширения связана с величиной $dn_G/d\lambda = -\lambda(d^2n/d\lambda^2)$.

В физической оптике и оптической связи несколько по-разному применяются также термины ***нормальная дисперсия*** и ***аномальная дисперсия***. В физической оптике в зависимости от знака производной $dn/d\lambda$ выделяют две спектральные области: область нормальной дисперсии, где фазовый показатель преломления убывает с ростом длины волны ($dn/d\lambda$<0), и область аномальной дисперсии, где фазовый показатель преломления возрастает с ростом длины волны ($dn/d\lambda$>0).

В оптической связи также выделяют две спектральные области в зависимости от знака производной группового показателя

преломления $dn_G / d\lambda = -\lambda(d^2n/d\lambda^2)$: область нормальной дисперсии, где групповой показатель преломления убывает с ростом длины волны ($dn_G / d\lambda < 0$), и область аномальной дисперсии, где групповой показатель преломления возрастает с ростом длины волны ($dn_G / d\lambda > 0$).

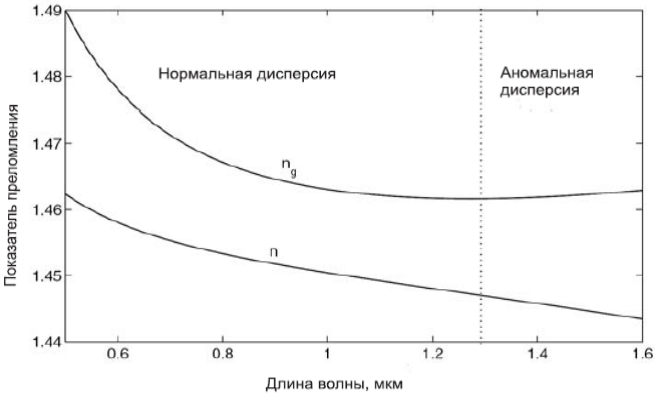

Рис.6.2. Спектральные зависимости фазового и группового показателей преломления плавленого кварца.

Распространение световых импульсов

Световые импульсы можно рассматривать как суперпозицию плоских монохроматических световых волн с различными частотами и волновыми векторами. Ограничимся рассмотрением одномерной задачи, т.е. задачи о распространении импульса, зависящего только от одной координаты, в однородной и изотропной среде можно ограничиться скалярным вариантом теории, понимая под $E(t,z)$ какую-либо декартову компоненту поля. Пренебрежем также поглощением света средой.

Рассмотрим задачу, когда задан световой импульс $E(t,0)$ в плоскости $z = 0$ и требуется установить характер эволюции этого импульса по мере распространения в пространстве, т.е. $E(t,z)$.

Введем обозначение

$$E(t,0) = A^T(t)\exp(-i\omega_0 t) \qquad (6.8)$$

где ω_0 – некоторая средняя частота, которую принято называть оптической несущей частотой, $A(t,0)$ – огибающая амплитуды

входного светового сигнала, в общем случае комплексная функция времени. Представим $A(t,0)$ в виде интеграла Фурье:

$$A^T(t,0) = \int_{-\infty}^{\infty} A_F^T(\omega)\exp\{-i(\omega-\omega_0)t\}d\omega \qquad (6.9)$$

Функция $A_F(\omega)$ есть преобразование Фурье от амплитуды входного импульса:

$$A_F^T(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} A^T(t)\exp\{i(\omega-\omega_0)t\}dt \qquad (6.10)$$

Поскольку каждая Фурье-компонента светового импульса в линейной среде распространяется независимо от других, то мы можем сконструировать функцию $E(t,z)$ в виде

$$E(t,z) = \int_{-\infty}^{\infty} A_F^T(\omega)\exp\{-i[\omega t - k(\omega)z]\}d\omega, \quad k^2(\omega) = \frac{\omega^2}{c^2}\varepsilon(\omega). \qquad (6.11)$$

Выражение (6.11) справедливо при любой спектральной плотности $A_F(\omega)$ и при любой зависимости волнового вектора от частоты $k(\omega)$, и в этом смысле оно общее и точное. Но, поскольку в данном решении пренебрегается нелинейными эффектами, то с этой точки зрения оно является приближенным.

Самый простой случай отвечает распространению импульса в вакууме:

$$E(t,z) = \int_{-\infty}^{\infty} A_F^T(\omega)\exp\left[-i\omega\left((t-\frac{z}{c}\right)\right]d\omega = E\left(t-\frac{z}{c},0\right), \quad k=\frac{\omega}{c}, \qquad (6.12)$$

т.е. сигналы (импульсы) распространяются как целое со скоростью света в вакууме и не изменяют своей формы. Это вполне понятно, т.к. гармоники распространяются с одинаковой скоростью, и между ними не возникает дополнительный набег фаз.

В реальных средах с дисперсией скорости распространения гармоник различны и между ними указанные набеги фаз возникают. Рассмотрим детальнее распространение импульса с относительно узким спектром $A_F(\omega)$ и центральной частотой ω_0. Разложим зависимость $k(\omega)$ в степенной ряд около ω_0 и ограничимся первыми членами разложения:

$$k(\omega) = k_0 + \frac{dk}{d\omega}(\omega-\omega_0) + \frac{1}{2}\frac{d^2k}{d\omega^2}(\omega-\omega_0)^2, \; k_0 \equiv k(\omega_0). \qquad (6.13)$$

Если пренебречь квадратичным членом разложения, что допустимо при выполнении условия

$$\frac{1}{2}\left|\frac{d^2k}{d\omega^2}\right|(\omega-\omega_0)^2 z << 1 \qquad (6.14)$$

и учесть, что согласно (6.1) $dk/d\omega = 1/v_G$, то (6.11) примет ид

$$E(t,z)=\exp\{-i(\omega_0 t-k_0 z)\}\times\int_{-\infty}^{\infty}A_F^T(\omega)\exp\{-i(t-z/v_G)(\omega-\omega_0)\}d\omega. \qquad (6.15)$$

Сопоставляя выражения (6.15) и (6.9), можно заключить, что интеграл в (6.15) совпадает с $A^T(t-z/v_G)$ и эволюция светового сигнала описывается выражением

$$E(t,z)=\exp\{-i(\omega_0 t-k_0 z)\}A^T(t-z/v_G). \qquad (6.16)$$

В рассматриваемом приближении как и в случае бигармонической волны быстро осциллирующая оптическая несущая $\exp\{-i(\omega_0 t-k_0 z)\}$ распространяется с фазовой скоростью $v_P=\omega_0/k_0$. Более медленно изменяющаяся огибающая $A(t-z/v_G)$ сохраняет свою форму и распространяется с групповой скоростью $v_G=d\omega/dk$.

Дифференцирование $k(\omega)$ по ω приводит к формуле, связывающей групповую и фазовую скорости

$$v_G=d\omega/dk=c\left(n+\frac{\omega dn}{d\omega}\right)^{-1}=v_P\left(1+\frac{\omega dn}{nd\omega}\right)^{-1}. \qquad (6.17)$$

В области прозрачности диэлектриков и полупроводников производная показателя преломления по частоте больше 0 и $v_G<v_P$, т.е. огибающая импульса отстает от несущей.

Волновое уравнение для огибающей светового импульса

Непосредственно из уравнений Максвелла можно получить уравнение, описывающее распространение световых сигналов в материальной среде. Уравнение (5.5) Лекции 5 с учетом (5.6) и равенства

$$\vec{\nabla}\times\vec{\nabla}\times\vec{E}\equiv\nabla(\nabla\cdot\vec{E})-\nabla^2\vec{E}=\nabla^2\vec{E} \qquad (6.18)$$

можно записать в виде:

$$\nabla^2\vec{E}-\frac{1}{c^2}\frac{\partial^2\vec{E}}{\partial t^2}=\mu_0\frac{\partial^2\vec{P}_L}{\partial t^2}+\frac{\partial^2\vec{P}_{NL}}{\partial t^2} \qquad (6.19)$$

Фурье-компоненты электрического поля удовлетворяют уравнению Гельмгольца (5.21). Из этого уравнения (в пренебрежении нелинейными эффектами и с учетом малости второй производной от медленно меняющейся огибающей) можно получить уравнение для Фурье компонент медленно меняющейся огибающей:

$$2ik_0\frac{\partial A_F}{\partial z}+(k^2-k_0^2)A_F=0 \qquad (6.20)$$

Воспользовавшись приближенным выражением

$$(k^2-k_0^2)\cong 2k_0(k-k_0) \qquad (6.21)$$

получим следующее уравнение

$$\frac{\partial A_F}{\partial z} = i[k(\omega) - k_0]A_F \qquad (6.22)$$

Физический смысл этого уравнения состоит в том, что каждая спектральная компонента огибающей импульса в процессе распространения приобретает фазовый сдвиг, величина которого зависит от частоты. Воспользовавшись разложением $k(\omega)$ в ряд Тейлора (6.13) и используя обратное преобразование Фурье уравнения (6.20) получим уравнение для A_F во временной области. Учитывая, что операция фурье-преобразования $(\omega-\omega_0)$ заменяется оператором дифференцирования $i(\partial/\partial t)$, получаем

$$\frac{\partial A}{\partial z} + \frac{dk}{d\omega}\frac{\partial A}{\partial t} + \frac{i}{2}\frac{d^2k}{d\omega^2}\frac{\partial^2 A}{\partial t^2} = 0. \qquad (6.23)$$

Для анализа эволюции световых импульсов и сигналов удобно использовать систему координат, движущуюся совместно с импульсом с групповой скоростью v_G (так называемые бегущие координаты). Переход к бегущим координатам осуществляется заменой переменных:

$$T = t - z/v_G \equiv t - z\,dk/d\omega \qquad (6.24)$$

В новых координатах уравнение (6.22) будет иметь следующий вид:

$$\frac{\partial A}{\partial z} + \frac{i}{2}\frac{d^2k}{d\omega^2}\frac{\partial^2 A}{\partial T^2} = 0. \qquad (6.25)$$

Это уравнение аналогично уравнению, описывающему дифракцию световых пучков в поперечном направлении в параксиальном приближении. Дисперсионное уравнение в точности совпадает с дифракционным в одномерном случае (т.е. при распространении световых пучков в планарных волноводах) при замене $dk/d\omega$ на $-\lambda/2\pi$, где λ – длина световой волны. В самом деле, временные эффекты, связанные с дисперсией, имеют близкие аналогии с пространственными дифракционными эффектами. Аналогия между пространственной дифракцией и временной дисперсией оказывается во многих случаях очень продуктивной.

Векторные волны

В общем случае решением уравнения Максвелла является волна, состоящая из нескольких компонент, т.е. имеющая векторный, а не скалярных характер. Рассмотрим простейшее решение системы уравнений в виде плоской волны. Каждая из компонент поля зависит от пространственных и временных переменных только через их комбинацию:

$$u = \vec{r}\vec{s} - vt, \qquad (6.26)$$

т.е. $\vec{E} = \vec{E}(\vec{r}\vec{s} - vt)$ (6.27)

$\vec{H} = \vec{H}(\vec{r}\vec{s} - vt)$ (6.28)

где $\vec{s}$ - единичный вектор в направлении распространения волны

$$\frac{\partial \vec{E}}{\partial t} = -v\frac{\partial \vec{E}}{\partial u} \tag{6.29}$$

$$\left(rot\vec{E}\right)_x = \frac{\partial E_z}{\partial \varphi} - \frac{\partial E_y}{\partial z} = E_z{}' s_y - E_y{}' s_z = \left(\vec{s} \times \vec{E'}\right)_x .$$

(6.30)

Уравнения Максвелла тогда примут вид:

$$\left.\begin{array}{l} \vec{s} \times \vec{H'} + \varepsilon_0 \varepsilon v \vec{E} = 0 \\ \vec{s} \times \vec{E'} - \mu_0 \mu v \vec{H} = 0 \end{array}\right\}. \tag{6.31}$$

Считая постоянную интегрирования равной нулю, т.е. пренебрегая постоянным полем, и учитывая, что $\frac{v}{c} = \frac{1}{\sqrt{\varepsilon\mu}}$, получим решения:

$$\vec{E} = -\sqrt{\frac{\mu_0 \mu}{\varepsilon_0 \varepsilon}}(\vec{s} \times \vec{H}); \tag{6.32}$$

$$\vec{H} = \sqrt{\frac{\varepsilon_0 \varepsilon}{\mu_0 \mu}}(\vec{s} \times \vec{E}). \tag{6.33}$$

Умножая скалярно полученные выражения на $\vec{s}$, получаем условие поперечности электромагнитной волны:

$$\vec{E}\vec{s} = \vec{H}\vec{s}. \tag{6.34}$$

Оно показывает, что электрический и магнитный векторы лежат в плоскости, перпендикулярной направлению распространения.

Поляризация электромагнитной волны

Рассмотрим плоскую гармоническую волну. Каждая из компонент меняется по закону косинуса, т.е.

$$a\cos(\tau + \delta) \tag{6.35}$$

где $\tau = \omega\left(t - \frac{\vec{r}\vec{s}}{v}\right) = \omega t - \vec{k}\vec{r}$ (6.36)

Пусть волна распространяется вдоль оси z. Тогда, вследствие поперечности электромагнитной волны, у неё будут только компоненты x и y (вектор $\vec{s}$ направлен вдоль оси z).

Рассмотрим кривую, которую описывает конец вектора $\vec{E}$ в произвольной точке пространства. Эта кривая является геометрическим местом точек, координаты которых равны:

$$\left.\begin{aligned} E_x &= a_1\cos(\tau+\delta_1)\\ E_y &= a_2\cos(\tau+\delta_2)\end{aligned}\right\} \qquad (6.37)$$

Преобразуем уравнения:

$$\left.\begin{aligned} \frac{E_x}{a_1} &= \cos\tau\cos\delta_1-\sin\tau\sin\delta_1\\ \frac{E_y}{a_2} &= \cos\tau\cos\delta_2-\sin\tau\sin\delta_2\end{aligned}\right\} \qquad (6.38)$$

Следовательно,

$$\left.\begin{aligned} \frac{E_x}{a_1}\sin\delta_2-\frac{E_y}{a_2}\sin\delta_1 &= \cos\tau\sin(\delta_2-\delta_1)\\ \frac{E_x}{a_1}\cos\delta_2-\frac{E_y}{a_2}\cos\delta_1 &= \sin\tau\sin(\delta_2-\delta_1)\end{aligned}\right\} \qquad (6.39)$$

Возведя в квадрат каждое из уравнений (5.44), и сложив их, получим следующее выражение:

$$\left(\frac{E_x}{a_1}\right)^2+\left(\frac{E_y}{a_2}\right)^2-2\frac{E_xE_y}{a_1a_2}\cos\delta=\sin^2\delta, \qquad (6.40)$$

где $\delta=\delta_2-\delta_1$

Уравнение (6.40) носит название канонического сечения. Геометрическое место точек концов вектора напряженности электрического поля в общем случае представляет эллипс, который вписан в прямоугольник со сторонами $2a_1$ и $2a_2$. В этом случае говорят, что волна эллиптически поляризована (рис. 6.3(а)).

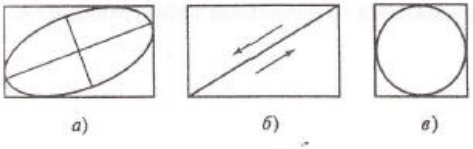

Рис. 6.3. Световая волна эллиптической поляризации при разных значениях δ*: (а)* $-\ 0<\delta<\frac{\pi}{2}$*; (б)* $-\ \delta=0$*; (в)* $-\ 0<\delta<\frac{\pi}{2}$, $a_1=a_2$.

В частном случае эллипс канонического сечения может выродиться либо в прямую линию (рис. 6.3(б)), либо в окружность (рис. 6.3(в)). В этих случаях имеет место линейная или круговая поляризация. Круговая поляризация, в зависимости от направления движения по окружности, может быть правой или левой.

Лекция 7. Оптические волноводы

Направляемые световые волны. Оптические волноводы. Волноводы на основе полного внутреннего отражения света. Моды планарного волновода. Дисперсионные кривые. Межмодовая дисперсия. Фотонно-кристаллические волокна.

Направляемые световые волны

В свободном пространстве свет распространяется прямолинейно, что было установлено опытным путем еще в древности и является основой геометрической оптики. Волновая теория указывает на то, что по мере прямолинейного распространения светового пучка происходит его дифракционное расширение. Для того чтобы направить световые волны по нужному пути необходимо видоизменить среду распространения таким образом, чтобы устранить дифракционную расходимость и обеспечить движение световых лучей по заданной траектории – вдоль определенных каналов распространения. Световые волны, распространяющиеся вдоль специальных каналов, называются ***направляемыми световыми волнами***, а сами каналы распространения – ***оптическими волноводами***.

Оптические волноводы

Для удержания световых волн в пределах некоторого канала могут быть использованы два явления: отражение от границы каналов (либо от граничной области конечной протяженности) или периодическая фокусировка лучей.

Первой попыткой реализации второго принципа была созданная фирмой IBM система периодически расположенных в пространстве линз. Однако реализация этой идеи в первоначальном виде оказалась коммерчески неперспективной и не получила развития. Тем не менее, можно считать, что принцип устранения дифракционного расширения световых пучков за счет рефракции реализован в современных градиентных диэлектрических волноводах.

Для реализации отражающих волноводов первоначально было предложено использовать полые металлические каналы круглого (трубки) или прямоугольного сечения. Такие металлические волноводы хорошо зарекомендовали себя для каналирования электромагнитных волн в другом диапазоне – диапазоне ВЧ и СВЧ радиоволн. Однако в оптическом диапазоне затухание в них оказалось

недопустимо большим и металлические волноводы не нашли практического применения.

Более продуктивным оказалось использование полного внутреннего отражения на границе раздела двух диэлектриков. В большинстве оптических волноводов: планарных, канальных, волоконных – используется явление полного внутреннего отражения света. Наиболее распространены кварцевые волоконные световоды, которые также называют оптическими волокнами (такое название является наиболее употребительным в прикладной технической литературе). Следует отметить, что для реализации планарных, канальных и других оптических волноводов наряду с диэлектриками могут использоваться полупроводниковые материалы в области прозрачности.

Сравнительно недавно начались интенсивные исследования оптических волноводов, использующих свойство нового класса оптических материалов – так называемых фотонных кристаллов с запрещенными фотонными зонами – отражать падающее на них световое излучение определенного спектрального диапазона длин волн. Физический механизм отражения света – резонансная брэгговская дифракция на периодической структуре. Такой механизм отражения и канализации света реализуется в планарных волноводах в фотонно-кристаллических материалах, в брэгговских оптических волокнах и фотонно-кристаллических волокнах с запрещенными фотонными зонами.

Волноводы на основе полного внутреннего отражения света

Если углы φ падения света на границу раздела двух сред таковы, что $\sin\varphi \geq n_2 / n_1$ (падающая волна распространяется в среде 1), то волна полностью отражается от границы раздела и не проникает в среду 2. На основе этого явления – полного внутреннего отражения (ПВО) действуют волоконные световоды, называемые также оптическими волокнами, а также планарные и канальные диэлектрические волноводы.

Волоконно-оптический волновод состоит из стеклянной сердцевины, окруженной оболочкой с меньшим значением показателя преломления. Сердцевина волокна, показанного на рис. 7.1, имеет постоянный показатель преломления n_1.

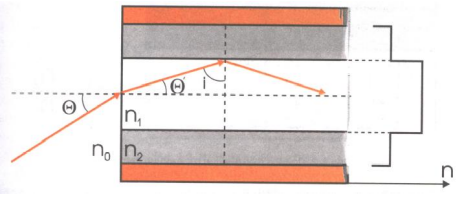

Рис. 7.1. Ход лучей в оптическом волокне со ступенчатым профилем показателя преломления

Показатель преломления скачком изменяется до значения n_2 на границе между сердцевиной и оболочкой. Зависимость показателя преломления от координаты вдоль диаметра волокна, которая называется профилем показателя преломления, в рассматриваемом волокне имеет вид прямоугольной ступеньки. Поэтому волокно такого типа называется волокном со ступенчатым профилем показателя преломления.

Если $n_2 < n_1$ волокно способно локализовать световой пучок в сердцевине за счет полного внутреннего отражения. При отражении от границы раздела двух сред световых лучей, падающих под углом i, меньше критического, преломленного луча не существует и световой луч отражается без затухания. Поэтому, без потерь распространяться внутри сердцевины ОВ будут только лучи, падающие на границу раздела сердцевины и оболочки под углом i, превышающим критический угол i_C.

Для того, чтобы определить какие из вводимых в волокно лучей будут им каналироваться найдем связь между углом θ падения луча на торец волновода и лучом i. Как следует из рис.7.1 они связаны соотношением:

$$n_0 \sin\theta = n_1 \sin\theta' = n_1 \cos i \qquad (7.1)$$

где n_0 — показатель преломления окружающей среды, n_1 - показатель преломления сердцевины. Критический угол определяется выражением

$$\sin i_C = n_1 / n_1 \qquad (7.2)$$

Используя два последних выражения и равенство $\cos^2 i_C = 1 - \sin^2 i_C$, получим значение максимального угла θ_m ввода для направляемых лучей

$$n_0 \sin\theta_m = (n_1^2 - n_2^2)^{1/2} \approx [2n_1(n_1 - n_2)]^{1/2}. \qquad (7.3)$$

Величина $n_0 \sin\theta_m$ называется числовой апертурой (NA) волновода. Для достижения высокой эффективности ввода волновод необходимо освещать источником с числовой апертурой, не

превышающей числовой апертуры волновода. Значение n_0 может быть равно 1, но, возможно, волновод приклеен к светодиоду или другому устройству клеем с показателем преломления близким к показателю преломления сердцевины.

У типичного многомодового волновода с n_1= 1.5 и $n_1 - n_2$=0.01 числовая апертура приблизительно равна 0.17. Когда n_0=1 это соответствует конусу входа световых лучей с половинным углом, равным 10°.

Из-за ограниченности диапазона углов распространения лучей вдоль волокна ввод излучения источника света в волновод не простая задача. Например, излучение светодиодов полностью заполняет полусферу, поэтому в волновод можно ввести только малую часть мощности излучения светодиода. Мощность, излучаемая в конус с половинным углом θ_m, пропорциональна $\sin^2\theta_m$. Поэтому, если сердцевина по крайней мере не меньше излучающей поверхности светодиода, эффективность ввода излучения светодиода в волновод равна $\sin^2\theta_m$ (Она меньше, если светодиод больше сердцевины). Для волновода из предыдущего примера эффективность ввода составляет примерно 3%.

В современных оптических волокнах распределение показателя преломления обычно существенно отличается от показанного на рис.7.1, но всегда в них можно выделить центральную часть – сердцевину с более высоким показателем преломления и оболочку с меньшим значением показателя преломления.

В интегральной оптике также используются оптические волноводы, основанные на явлении ПВО. К таким волноводам относятся планарные волноводы, часто имеющие несимметричное распределение показателя преломления, а также канальные волноводы различного типа.

Моды планарного волновода

Рассмотрим распространение световой волны в планарном волноводе. В общем случае такой волновод состоит из трех областей (оболочки, пленки и подложки) с показателями преломления оболочки n_c, пленки n_f и подложки n_s, которые отделены друг от друга планарными (плоскими границами) перпендикулярными оси x. Ось z является направлением распространения света. Предположим, что $n_f < n_s < n_c$, и что плоскость x=0 соответствует границе между оболочкой и пленкой. Следовательно, если толщина пленки равняется

d, поверхность границы раздела двух сред пленки и подложки расположена на плоскости x=d.

Направляемые ТЕ моды. Хотя планарные волноводы со ступенчатым распределением показателя преломления - это в сущности неоднородные структуры Внутри каждой из трех сред показатель преломления является постоянной величиной. Таким образом, рассматривая каждую среду по отдельности, можно записать волновое уравнение для ТЕ мод в следующем виде:

$$\frac{d^2 E_y}{dx^2} + \left[k_0^2 n^2 - \beta^2\right] E_y = 0. \qquad (7.3)$$

Будем искать направляемые моды, экспоненциально убывающие в оболочке и подложке. Константа распространения β таких мод должна удовлетворять условию:

$$k_0 n_s < \beta < k_0 n_f \qquad (7.4)$$

Если ввести эффективный показатель преломления направляемой моды N, то для него должно выполняться условие:

$$n_s < N < n_f \qquad (7.5)$$

С учетом сказанного, в каждой однородной области волновые уравнения (7.3) могут быть записаны в виде:

$$\frac{d^2 E_y}{dx^2} - \gamma_c^2 E_y = 0 \qquad x \geq 0 \quad \text{(оболочка)} \qquad (7.6)$$

$$\frac{d^2 E_y}{dx^2} + \kappa_f^2 E_y = 0 \qquad 0 > x > -d \quad \text{(пленка)} \qquad (7.7)$$

$$\frac{d^2 E_y}{dx^2} - \gamma_s^2 E_y = 0 \qquad x \leq -d \qquad \text{(подложка)} \qquad (7.8)$$

где введены параметры γc, κf, γs :

$$\gamma_c^2 = \beta^2 - k_0^2 n_c^2 \qquad (7.9)$$

$$\kappa_f^2 = k_0^2 n_f^2 - \beta^2 \qquad (7.10)$$

$$\gamma_s^2 = \beta^2 - k_0^2 n_s^2 \qquad (7.11)$$

Причем введенные параметры γc, κf, γs для направляемых мод являются реальными величинами. Пусть γc, κf, γs будут положительными. При решении системы дифференциальных уравнений (7.6) - (7.8) электрическое поле в оболочке, пленке и подложке может быть выражено следующим образом:

$$E_y = \begin{cases} Ae^{-\gamma_c x} & x \geq 0 \\ Be^{i\kappa_f x} + Ce^{-i\kappa_f x} & -d < x < 0 \\ De^{\gamma_s x} & x \leq -d \end{cases} \qquad (7.12)$$

Электрическое поле в оболочке также допускает дополнительные решения в форме $A' e^{+\gamma_c x}$, но т.к. увеличение экспоненциальной функции при x > 0 не имеет физического смысла для направляемых мод, то следует положить A' = 0. Путем аналогичных рассуждений устраняется и член D' exp (-γs x), соответствующий области подложки.

Граничные условия накладывают условия непрерывности на поверхности раздела двух сред оболочка - пленка (x = 0) и на границе пленка - подложка (x = -d) величин Ey и dEy/dx. Из граничных условий можно получить четыре уравнения, представляющие собой отношения величин A, B, C, D и β. Следовательно, у нас есть пять неизвестных величин, которые нужно определить из четырех уравнений. Действительно, один из постоянных параметров не может быть определен и будет оставаться независимым (например, параметр A). Квадрат этого параметра определяет мощность световой волны, переносимую данной модой. Решая эту систему уравнений, после громоздких но несложных вычислений, получается следующее уравнение:

$$\tan(\kappa_f d) = \frac{\frac{\gamma_c}{\kappa_f} + \frac{\gamma_s}{\kappa_f}}{1 - \frac{\gamma_c}{\kappa_f}\frac{\gamma_s}{\kappa_f}} \qquad (7.13)$$

Полученное выражение может быть рассмотрено как дисперсионное уравнение для ассиметричного планарного волновода со ступенчатым распределением показателя преломления. Данное уравнение является трансцендентным, включает в себя параметры, определяющие волноводную структуру (n_c, n_f, n_s и d), рабочую длину волны λ и постоянную распространения β направляемой моды. Из него можно численно вычислить постоянную распространения β. Кроме того, т.к. функция тангенциальная, то:

$$\tan(\kappa_f d) = \tan(\kappa_f d + m\pi), \qquad m = 0, 1, 2, \ldots \qquad (7.14)$$

и в общем случае существует несколько решений для константы распространения β, зависящих от целого числа m. Это целое число m называется порядком моды, а связанная с такой модой постоянная распространения обозначается так β_m.

Введем набор параметров - т.н. нормализованных параметров, для того, чтобы трансцендентное уравнение (7.14) было универсальным для любого ассиметричного волновода со ступенчатым распределением показателя преломления. Определим их следующим образом:

$$b = \left(N^2 - n_s^2\right) / \left(n_f^2 - n_s^2\right) \tag{7.15}$$

нормированный модовый показатель преломления

$$V = k_0 d \left(n_f^2 - n_s^2\right)^{1/2} \tag{7.16}$$

нормированная толщина пленки

$$a = \left(n_s^2 - n_c^2\right) / \left(n_f^2 - n_s^2\right) \tag{7.17}$$

показатель асимметрии

Т.к. величина эффективного показателя преломления, соответствующего ограниченной моде, лежит в диапазоне $n_s < N < n_f$, то нормализованный модовый показатель преломления b находится в области 0<b<1. С другой стороны, как следует из уравнения (7.16), нормализованная толщина пленки V напрямую связана с отношением толщины сердцевины (пленки) волновода к рабочей длине волны, т.е. $V \propto d/\lambda$. И наконец, показатель ассиметрии является нулевым в случае симметрических волноводов, и увеличивается по мере увеличения разности между показателями преломления оболочки и подложки.

Трансцендентное уравнение (7.13) принимает вид:

$$\tan\left[V\sqrt{1-b}\right] = \frac{\sqrt{\frac{b}{1-b}} + \sqrt{\frac{b+a}{1-b}}}{1 - \frac{\sqrt{b(b+a)}}{1-b}} \tag{7.18}$$

В общем случае, уравнение (7.18) (или (7.13)) имеет ограниченное число решений для ограниченного числа целых m, и т.о. волновод будет поддерживать ограниченное число каналируемых мод. В случае, когда дисперсионное уравнение имеет только одно решение (для m=0), волновод называется одномодовым. Кроме того, возможно, что некоторая структура не имеет решения трансцендентного уравнения (7.18), и в этом случае (для некоторых рабочих длин волн) волновод не поддерживает никаких каналируемых мод. В случае, когда дисперсионное уравнение имеет больше одного решения с разными m=0), волновод называется одномодовым.

Как только вычисленна постоянная распространения моды β (или b), просто определяются коэффициенты γ_c, κ_f и γ_s. Т.о. электрическое поле во всех трех областях теперь можно полностью определить:

$$E_y(x)=\begin{cases} A e^{-\gamma_c x} & x \geq 0 \\ A\left(\cos\kappa_f x-\dfrac{\gamma_c}{\kappa_f}\sin\kappa_f x\right) & 0 \leq x \leq -d \\ A\left(\cos\kappa_f d+\dfrac{\gamma_c}{\kappa_f}\sin\kappa_f d\right)e^{\gamma_s(x+d)} & x \leq -d \end{cases} \qquad (7.19)$$

Согласно данному выражению электрическое поле экспоненциально уменьшается в оболочке и в подложке, в то время как в пленке оно подчиняется синусоидальному закону как и ожидалось для поведения ограниченных мод. На рис. 7.2 показаны профили электрического поля для четырех ограниченных мод (m=0,1,2,3), поддерживаемые планарным волноводом, сердцевина которого имеет толщину 3 мкм и показатель преломления 1.50. С одной стороны сердцевина граничит с воздухом, а с другой – с подложкой с показателем преломления 1.43. Моды рассчитаны для длины волны λ=0.633 мкм. Из рис. 7.2 видно, что электрическое поле , а также его производная непрерывны на обоих границах раздела сред. Решение для E_y полностью определено, с точностью до постоянной A, квадрат которой пропорционален энергии, переносимой модой. Отметим, что число m, которое характеризует порядок моды совпадает с числом нулей функции профиля электрического поля.

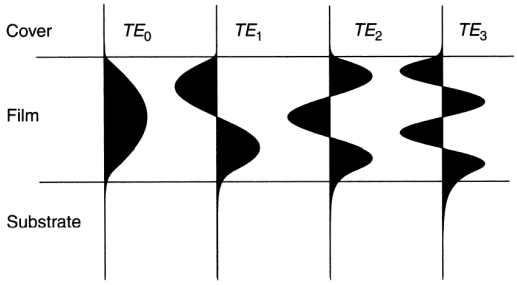

Рис. 7.2. Профили электрического поля направляемых TE мод планарного асимметричного диэлектрического волновода. Параметры волновода: d= 3 мкм; n_f *=1.50,* n_c *=1 (воздух); и* n_s *=1.43. Длина световой волны* λ *=0.633 мкм.*

Дисперсионные кривые

На рис. 7.3 показано численное решение дисперсионного уравнения для симметричного волновода. Решение уравнения

представлено в виде зависимости параметра b от V для мод следующих порядков: m=0, m=1, m=2 и m=3. Зная величину обобщенного параметра V можно определить число направляемых мод. Например, симметричный волновод, характеризуемый V=4, будет поддерживать две ТЕ моды (m=0,1).

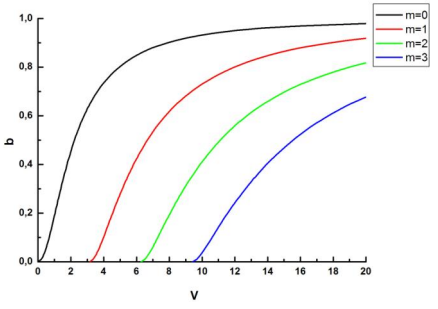

Рис. 7.3. Зависимость нормированного показателя преломления моды $b = \left(N^2 - n_s^2\right) / \left(n_f^2 - n_s^2\right)$ *от нормированной толщины симметричного планарного волновода* $V = k_0 d \left(n_f^2 - n_s^2\right)^{1/2}$.

Из дисперсионных кривых можно рассчитать зависимость фазовой и групповой скоростей от волнового вектора

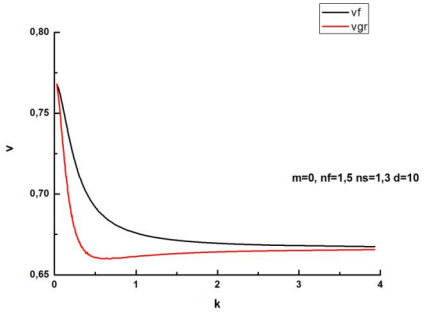

Рис. 7.4. Зависимости фазовой и групповой скоростей от волнового вектора для основной моды.

Межмодовая дисперсия

Каждая мода распространения характеризуется частотой световой волны ω, определенным пространственным распределением

огибающей электрического поля в поперечном сечении и постоянной распространения β вдоль оптической оси. При этом фазовая скорость определяется выражением:

$$V_F = \omega / \beta , \tag{7.20}$$

а групповая:

$$V_{GR} = \partial\omega / \partial\beta . \tag{7.21}$$

Если в волноводе распространяется несколько направляемых мод, то распределение поля в поперечном сечении есть интерференционная сумма полей мод:

$$E_\Sigma = \sum E_j \tag{7.22}$$

Поскольку моды ортогональны, то суммарная мощность также равна сумме мощностей световых импульсов, переносимых отдельными модами (парциальными импульсами).

$$P_\Sigma = \sum P_j \tag{7.23}$$

Входной световой импульс с огибающей мощности $P_\Sigma(t,0) = P_\Sigma^0 U_\Sigma^{IN}(t)$ складывается из парциальных импульсов поперечных мод $P_j(t,0) = P_j^0 U_j^{IN}(t)$.

Если пренебречь в первом приближении хроматической дисперсией, то форма огибающей парциальных импульсов не изменяется. Однако, из-за различия групповых скоростей распространения различных мод временные задержки выходных импульсов разных мод относительного входного импульса оказываются различными. Это приводит к расширению импульса при его распространении в волноводе.

Фотонно-кристаллические волокна

Фотонно-кристаллические оптические волокна – это световоды нового типа, отличающиеся по своей архитектуре, принципу действия и свойствам от обычных оптических волокон. В общем случае они представляют собой микроструктуру с периодически или апериодически промодулированным показателем преломления оболочки. В большинстве случаев для ее создания используют стекло или кварц с воздушными отверстиями. Однако существуют и другие типы волокон, например полимерные фотонно-кристаллические волокна и фотонно-кристаллические волокна, структура которых содержит стекло двух видов.

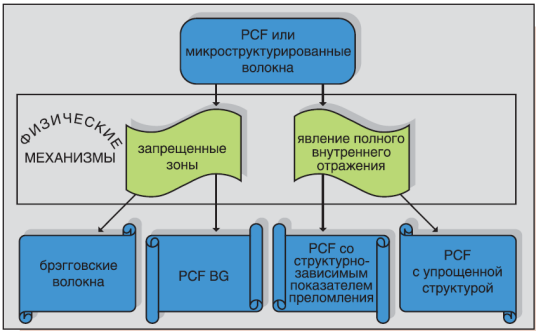

Рис. 7.5. Классификация фотонно-кристаллических волокон

Все многообразие созданных фотонно-кристаллических волокон можно разделить на два больших класса: волокна, в основе работы которых лежит эффект брэгговского отражения, и волокна, каналирование света в которых определяется эффектом полного внутреннего отражения (см. рис.7.5). Различные виды фотонно-кристаллических волокон показаны на рис.7.6.

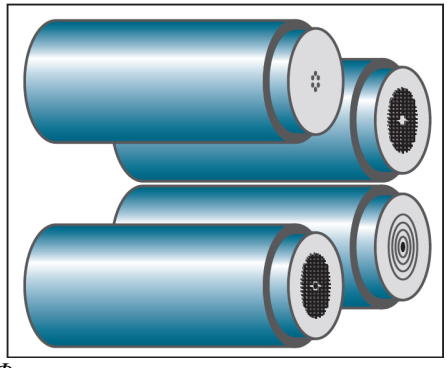

Рис. 7.6. Фотонно-кристаллические волокна различного типа:с упрощенной структурой и с периодической структурой на эффекте полного внутреннего отражения); брэгговские волокна и фотонно-кристаллические с запрещенной фотонной зоной (PCF BG)

Фотонно-кристаллические волокна, использующие эффект полного внутреннего отражения, содержат сердцевину с показателем преломления превышающим эффективный показатель преломления

оболочки, для изменения которого и создаются микроструктуры различного типа.

К фотонно-кристаллическим волокнам, использующим эффект брэгговского отражения, относятся брэгговские волокна и фотонно-кристаллические волокна с запрещенными фотонными зонами (PCF BG). Брэгговские волокна - это полые волноводы с отражающими стенками, выполненными из многослойного диэлектрического покрытия. Фотонно-кристаллические волокна с запрещенными фотонными зонами – это волокна, оболочка которых представляет собой двумерный фотонный кристалл. То есть оболочка в поперечном сечении обладает двумерной периодической структурой.

Один из способов изготовления фотонно-кристаллических волокон заключается в вытяжке при высокой температуре из преформы, набранной из полых волокон. Структура таких волокон показана на рис.7.7.

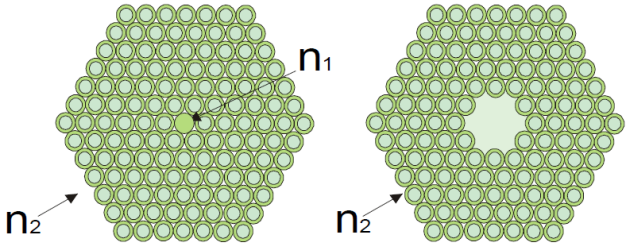

Рис. 7.7. Типы фотонно-кристаллических волокон с периодической структурой оболочки: а – фотонно-кристаллическое волокно с полным внутренним отражением ($n_2 < n_1$,); б – полое фотонно-кристаллическое волокно с фотонными запрещенными зонами

Уникальность микроструктурированных волокон для лазерной физики, нелинейной оптики и оптических технологий обусловлена возможностью управления дисперсией волноводных мод. Управление дисперсионными свойствами волноводных мод открывает новые возможности в области оптических телекоммуникаций и в оптике сверхкоротких импульсов. Имеется целый ряд режимов, характерных только для нелинейно-оптических взаимодействий в фотонно-кристаллических волокнах, не наблюдающихся ни в объемных твердых телах, ни в обычных оптических волокнах.

Лекция 8. Волоконно-оптическая связь

Развитие волоконно-оптической связи. Структура и основные компоненты ВОЛС. Принцип работы и виды WDM систем. Оптические мультиплексоры WDM

Развитие волоконно-оптической связи

В настоящее время волоконно-оптические линии связи (ВОЛС) используются в сетях практически всех масштабов: корпоративных сетях и сетях доступа, городских и региональных сетях, междугородных и трансконтинентальных линиях связи. И чем больше протяженность, чем выше скорость передачи, тем более заметны преимущества технологии ВОЛС по сравнению с другими. Анализ развития протяженных линий связи показывает, что в качестве среды передачи информации нет никакой альтернативы волокну. В течение последних лет наблюдается соревнование одной волоконнооптической технологии с новой, более совершенной волоконно-оптической технологией. Основные этапы эволюции протяженных ВОЛС представлены на рис.8.1.

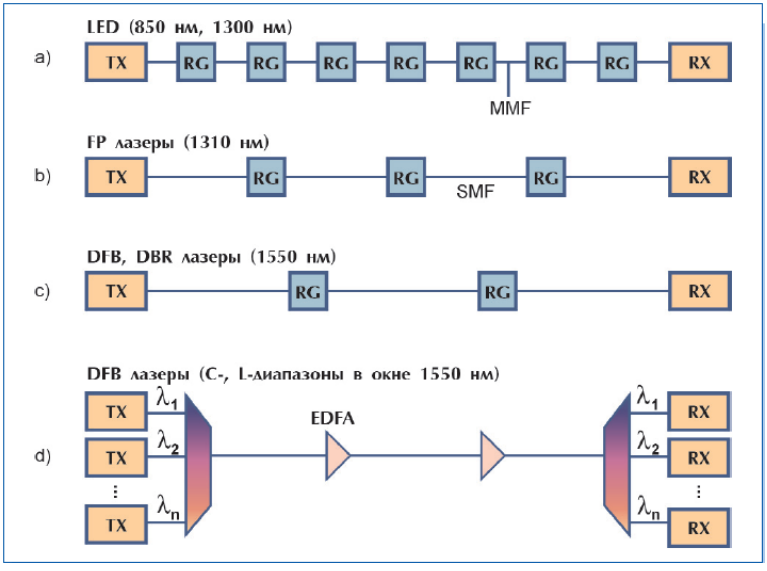

Рис. 8.1. Эволюция волоконно-оптических линий связи

Первую волоконно_оптическую коммуникационную систему компания Standard Telephones and Cables построила в сентябре 1975 г. В 1977 г. сразу несколько компаний сделали независимые заявления о том, что телефонный трафик был передан в реальном времени по оптическому волокну в их испытательных сетях. Это были компании AT&T, General Telephone and Electronics, British Post Office и другие. Системы с многомодовыми волокнами (MMF) составляли основу протяженных ВОЛС того времени. Наряду с градиентным многомодовым волокном (волокно G.651) в их состав входили светодиоды на основе арсенида галлия, излучающие на длине волны 850 нм. Поскольку потери в волокне на этой длине волны были более чем существенны (3 дБ/км), такие линии связи строились с большим числом близко расположенных друг к другу регенераторов. Эти оптические магистрали были наземными, а для межконтинентальной связи все еще использовались подводные коаксиальные кабели.

С появлением одномодового волокна – стандартного одномодового волокна или волокна G.652, – стало ясно, что значительно перспективней вести передачу на длине волны 1300 нм – меньше потери и дисперсия. Использование одномодового волокна позволяет передавать оптические сигналы с большей скоростью и на большие расстояния.

Сначала сложно было реализовать на практике преимущество нового типа волокна. Однако развитие технологии производства и улучшение технологии сварки одномодового волокна, серийное производство лазеров на длине волны 1300 нм способствовали быстрому моральному старению протяженных систем связи на основе многомодового волокна. Коммерческие преимущества новой технологии продемонстрировала компания MCI в 1982г. Система связи компании MCI, функционирующая на длине волны 1300 нм, передавала данные на 50% быстрее, чем система AT&T, использовавшая градиентное многомодовое волокно и передачу на двух длинах волн, 850 и 1300 нм. При этом MCI смогла увеличить расстояние между повторителями с 7 до 30 км! Это показало огромное превосходство одномодового волокна перед многомодовым градиентным для наземных протяженных линий связи. Другие компании, специализирующиеся на строительстве протяженных линий связи, быстро сделали выбор в пользу новой технологии на основе одномодового волокна.

Между тем трансатлантические кабельные операторы продолжали терять рынок – подводные коаксиальные кабельные системы не могли больше противостоять появляющимся системам спутниковой связи –

и в поисках ресурсов для увеличения полосы пропускания вынуждены были рассмотреть возможность использовать волокно. Подводные оптические кабели (ОК) начали производить позднее, чем кабели для наземных волоконно-оптических магистралей. Это было обусловлено сложностью изготовления подводного ОК – нагрузки на кабель и соответственно на волокна при укладке кабеля в грунт значительно меньше. Тем не менее в 1988 г. после нескольких лет планирования и строительства консорциум компаний, ведомый AT&T, сдал в эксплуатацию первую трансатлантическую волоконно-оптическую систему TAT_8, состоящую из 3 пар одномодовых волокон, работающую на длине волны 1300 нм с расстоянием между повторителями 60 км.

Минимальные значение потерь в стандартном одномодовом волокне 0,2–0,25 дБ/км достигается на длине волны, близкой к 1550 нм. Минимальная хроматическая дисперсия, в окрестности нуля, достигается на длине волны 1310 нм. Чтобы обеспечить высокую скорость передачи на большие расстояния, необходимо свести к минимуму потери и дисперсию, причем на одной и той же длине волны.

Прямолинейным решением было создание волокна со смещенной дисперсией (DSF, волокно G.653). Это волокно, имеющее нулевую дисперсию в окрестности длины волны 1550 нм, обещало быть очень привлекательным для одноканальной передачи. Однако две появившиеся впоследствии технологии – DWDM и EDFA – показали несостоятельность волокна DSF. Четырехволновое смешение, эффект, выражающийся в появлении дополнительных паразитных сигналов на частотах, являющихся комбинацией рабочих частот, которые также усиливаются, проходя через каскады усилителей EDFA. Этот эффект становится заметным при многоволновой передаче.

В 1994 г. создается волокно с ненулевой смещенной дисперсией (NZDSF, волокно G.655), в котором длина волны нулевой дисперсии вынесена за пределы зоны усиления EDFA. И это волокно также оказалось не способным удовлетворить растущие потребности. Рост числа каналов, канальной скорости передачи, увеличение протяженности сегментов между усилителями – все эти факторы по отдельности и тем более вместе требуют увеличения мощности излучения, вводимого в волокно.

В 1998 г. разрабатываются еще более совершенные волокна NZDSF с увеличенной эффективной площадью поперечного сечения. Поскольку проявление нелинейных эффектов зависит от плотности световой энергии в волокне, то увеличение эффективной площади

приводит к ослаблению влияния нелинейных эффектов и увеличивает дальность и скорость передачи информации.

Дальнейшие этапы развития ВОЛС связаны с внедрением эрбиевых оптических усилителей и технологии спектрального мультиплексирования. В настоящее время идут работы по увеличению скорости передачи информации по каждому спектральному каналу. При этом рассматриваются возможности использования новых, многоуровневых форматов модуляции.

За последние 30 лет оптическая связь кардинально изменила информационную структуру мира. Спрос на емкость каналов связи, а в терминологии IT-специалистов – на полосу пропускания удваивается приблизительно за несколько лет. Единственное средство, способное удовлетворить столь быстро растущие потребности в объемах передаваемой информации – волоконно-оптические сети связи (ВОСС). Внедрение эффективных полупроводниковых лазеров, оптического волокна с малыми потерями, эрбиевых усилителей и технологии спектрального уплотнения (WDM) обеспечили возможность передачи десятков и даже сотен каналов с разными длинами волн по одному волокну с суммарной скоростью, превышающей 10 Тбит/с.

Земной шар сегодня окутан сетью оптических волокон, многократно пересекающих океаны и континенты и соединяющих города по всему миру. Грандиозная общественная сеть связи играет важнейшую роль в современном обществе, обеспечивая эффективную передачу всех видов цифровой информации: страниц Интернета, электронных писем, музыки в формате mp3 и изображений или традиционных разговоров и факсов. Кроме глобальной общественной сети связи существует множество частных и специализированных сетей.

Глобальная общественная сеть связи, объединяющая сети связи разных стран в единую сеть, имеет очень сложное строение, не подчиняющееся ни строгой иерархической, ни функциональной структуризации. Она состоит из множества так называемых общественных сетей, работа которых поддерживается операторами связи. Операторы связи (иногда их называют провайдерами) предоставляют разнообразные услуги связи заказчикам (абонентам).

Основу общественных сетей связи составляют телефонные сети связи. Они организованы иерархически. Современная телефонная сеть представляет собой многоуровневую, частично древовидную сеть цифровых концентраторов и коммутаторов, связанных между собой цифровыми каналами. Только на «последней миле» абоненты

телефонных сетей связаны с телефонными станциями низкоскоростными каналами связи. Такие сети называются сетями доступа.

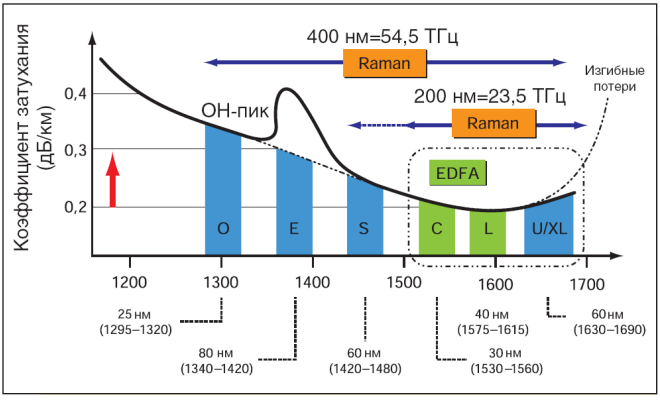

Рис. 8.2. Системы дальней связи в ближайшее10-20 лет должны освоить всю доступную для передачи сигналов полосу оптического волокна от 1300 нм до 1700 нм

Возможности эксплуатируемых сегодня систем связи будут ограничивать скорость развития трансокеанских систем связи уже начиная с 2010 года. Поэтому настало время поиска новых физических идей для реализации систем связи следующего поколения. Такие системы связи должны будут обеспечить к 2025 году возможность использования всего доступного оптического диапазона кварцевого волокна, равного 400 нм, или примерно 55 ТГц, при спектральной эффективности 10 бит/Гц.

Структура и основные компоненты ВОЛС

Простейшая волоконно-оптическая система связи передает информацию между двумя точками. Такие системы связи точка-точка называют волоконно-оптическими линиями связи (ВОЛС). В состав ВОЛС входят:

- передатчик – устройство, преобразующее входные управляющие электрические сигналы в выходные световые сигналы;
- приемник – устройство, преобразующее входные оптические сигналы в выходные электрические сигналы;

- физическая среда передачи информационных сигналов – оптическое волокно;
- регенераторы и/или оптические усилители.

Как правило, источниками световых сигналов служат полупроводниковые лазеры или светодиоды. Световые сигналы, выходящие из передатчика, вводятся в снабженное разъемом волокно и передаются по волоконно-оптической линии. В конце линии свет поступает в фотоприемник, преобразующий его в электрические сигналы, которые затем обрабатываются и используются в приемном оборудовании. Таким образом, обязательными элементами ВОЛС являются передатчик, оптическое волокно и приемник. Для увеличения дальности передачи информации используются регенераторы или оптические усилители сигналов. Схема ВОЛС, используемой для передачи информации на большое расстояние, показана на рис.8.3.

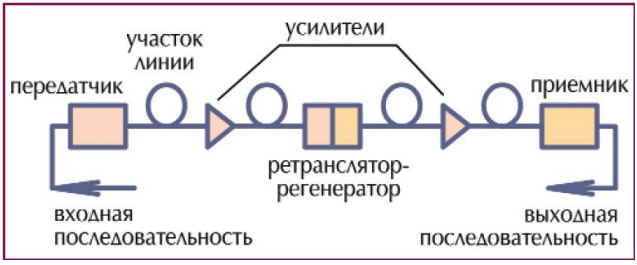

Рис.8.3. Структура одноканальной волоконно-оптической линии связи

В оптических передатчиках используются полупроводниковые лазеры и в системах передачи с небольшой скоростью светодиоды. Полупроводниковые лазеры на GaAlAs работают в диапазоне длин волн от 700 нм до 1000 нм, а на InGaAsP – от 1000 нм до 1700 нм. В редких случаях используются передатчики на основе волоконных лазеров.

В оптических приемниках используются полупроводниковые фотодиоды. Различные полупроводниковые материалы для фотодетекторов эффективно работают в различных спектральных областях: Si используется в диапазоне от 650 нм до примерно 950 нм, InGaAsP – от 950 нм до 1150 нм, Ge – примерно от 1100 нм до 1550 нм, а InGaAs – от 1300 нм до 1700 нм. Таким образом, существующие полупроводниковые передатчики и приемники работают в широком диапазоне длин волн, представляющих интерес для оптической связи.

Оптические усилители предназначены, в основном, для работы в спектральных областях, используемых в одномодовом волокне. Усилители на волокне, легированном эрбием (EDFA), работают в области 1545 нм и являются в настоящее время основным типом оптических усилителей.

Реже используются усилители на фтористом волокне, легированном празеодимом (PDFFA), работающие в области 1305 нм. Однако по эффективности и другим рабочим характеристикам празеодимовые усилители значительно уступают эрбиевым. Поэтому задача создания эффективных усилителей вне области работы эрбиевых усилителей не решена до настоящего времени.

Полупроводниковые оптические усилители (SOA) и малошумящие ВКР-усилители, которые в переводной литературе называются рамановскими волоконными усилителями (RFA) могут охватить спектральный диапазон от волн ниже 1300 нм до волн выше 1600 нм (см. рис.1). Для некоторых приложений используются комбинации различных типов оптических усилителей, чтобы обеспечить широкую и ровную спектральную полосу усиления при малом уровне шумов.

В международных стандартах определены спектральные диапазоны, в которых предполагается использование одномодового кварцевого оптического волокна. Краткая информация о них приведена в Таблице 8.1.

Таблица 8.1. Одномодовые спектральные диапазоны.

Диапазон	Описатель	Область [нм]
Диапазон O	Исходный	от 1260 до 1360
Диапазон E	Расширенный	от 1360 до 1460
Диапазон S	Коротких волн	от 1460 до 1530
Диапазон C	Условный	от 1530 до 1565
Диапазон L	Длинных волн	от 1565 до 1625
Диапазон U	Сверхдлинных волн	от 1625 до 1675

Диапазон сверхдлинных волн U предназначен только для целей возможного технического обслуживания и не предназначен для целей передачи сигналов, переносящих трафик.

Ожидается, что в ближайшем будущем различные приложения, с оптическими усилителями и без них, будут использовать передачу сигналов по одномодовым оптическим волокнам во всей области от 1260 нм до 1625 нм.

В многомодовых кварцевых волокнах используются два диапазона: от 770 до 910 нм и от 1270 до 1380 нм.

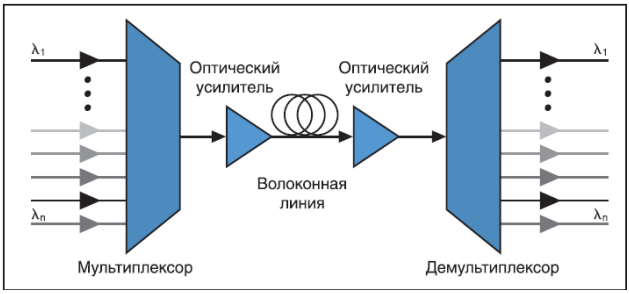

Рис.8.4. Структура волоконно-оптической линии связи со спектральным мультиплексированием

Принцип работы и виды WDM систем

Принцип работы WDM-систем поясняет рис. 8.4. Световые сигналы с разными длинами волн, генерируемые несколькими оптическими передатчиками, объединяются мультиплексором и вводятся в оптическое волокно линии связи. При больших расстояниях передачи на линии связи устанавливается один или несколько оптических усилителей. На приемном конце линии связи демультиплексор принимает составной сигнал, выделяет из него исходные компоненты с разными длинами волн и направляет их на соответствующие фотоприемники.

Такая система передачи «точка-точка» обеспечивает увеличение пропускной способности линии связи между двумя узлами. Однако возможности и преимущества технологии WDM в еще большей степени раскрываются в сложных насыщенных сетях связи, содержащих много различных узлов. На промежуточных узлах некоторые каналы могут быть добавлены или выделены из составного сигнала посредством мультиплексоров ввода/вывода, а остальные каналы проходят через узел без преобразования в электрический сигнал. В некоторых узлах устройства оптической кросс-коммутации позволяют перенаправлять каналы по новым направлениям (*рис. 8.5*).

Первые исследования WDM-технологии, проведенные в 1980-х годах, продемонстрировали возможность объединения оптических несущих, разделенных спектральным интервалом 10–25 нм, для передачи сигналов по многомодовому волокну в локальных сетях, при этом рабочие длины волн лежали в первом (850 нм) и втором (1310 нм) окнах прозрачности. Однако эти работы не привели к разработке промышленных систем, главным образом, по экономическим соображениям.

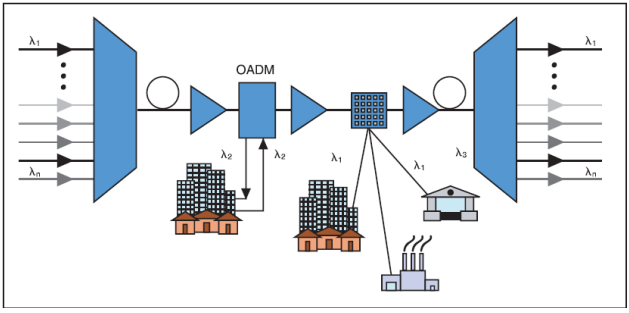

Рис.8.5. Принцип работы WDM-систем передачи информации в сложных сетях. OADM – мультиплексор ввода/вывода, ОС – оптический кросс-коммутатор

Первыми WDM-системами, нашедшими практическое применение, стали двухволновые WDM-системы, объединившие две основные несущие длины волн 1310 нм и 1550 нм в одном одномодовом волокне. Практический успех двухволновых WDM-систем обусловлен тем, что они позволяют либо удвоить скорость передачи сигналов по одному волокну, либо создать дуплексные системы на одном волокне, не изменяя существующего активного оборудования и используя простые и надежные двухволновые мультиплексоры.

Простота таких систем, обусловленная очень большим спектральным интервалом (более 200 нм), вместе с тем ограничивает дальнейший рост их пропускной способности. Реально двухволновые WDM-системы позволяют подключить еще только один канал с длиной волны 1650 нм или 1490 нм. Двухволновые WDM-системы широко используются в сетях доступа, в частности, в пассивных оптических сетях (PON).

В середине 1990-х годов благодаря широкому внедрению оптических усилителей на основе волокон, легированных эрбием (EDFA, Erbium doped fiber amplifier), начинает бурно развиваться технология спектрального мультиплексирования с плотным расположением спектральных каналов, для обозначения которой используется аббревиатура DWDM (Dense WDM). Экономическая эффективность систем DWDM в системах дальней связи резко увеличилась с применением оптических усилителей, так как одно устройство – усилитель – заменило десятки регенераторов, использовавшихся до появления оптических усилителей для

восстановления оптических сигналов каждого спектрального канала отдельно. Системы электрической регенерации сигналов, применяемые, например, в сетях SDH, являются весьма дорогими и, кроме того, протокольно зависимыми, так как они могут воспринимать только определенный вид кодирования сигнала. В силу того, что основной рабочий диапазон усилителей EDFA лежит в пределах длин волн 1525–1565 нм, появилась необходимость вместить в этот промежуток как можно больше каналов. Наиболее широкое распространение получили системы, в которых предусмотрено расположение каналов с частотным интервалом 100 ГГц, что в области 1550 нм соответствует спектральному интервалу 0,8 нм. Ведутся работы по созданию систем с частотным интервалом 50 ГГц (0,4 нм) и даже 25 и 12,5 ГГц.

Технология DWDM оказалась незаменимой в линиях дальней связи, в которых необходимо передавать огромные потоки информации на большие расстояния, требующие применения оптических усилителей. Кроме того, в последнее время активно развиваются городские сети и сети доступа, в которых также целесообразно применение технологий спектрального мультиплексирования. В некоторых из них не требуются столь высокие суммарные потоки информации, которые обеспечивает технология DWDM. Поэтому вновь возродился интерес к WDM-системам с менее плотным расположением спектральных каналов. Такие системы называются системами с грубым спектральным мультиплексированием, и для них принято международное обозначение CWDM (Coarse WDM). Международным стандартом ITU G.694.2 установлена спектральная сетка для центральных длин волн CWDM-каналов. Соседние каналы разделены спектральным интервалом 20 нм в диапазоне длин волн от 1270 до 1610 нм. Стандарт определяет и область применения технологии CWDM – городские сети с расстоянием до 50 км. Основное преимущество технологии CWDM перед технологией DWDM – более простые компоненты и следовательно меньшая стоимость.

Главный недостаток технологии CWDM заключается в ограниченных возможностях масштабирования т.е. увеличения суммарного по всем каналам потока передаваемой информации по мере роста потребностей заказчика. Наибольшее количество спектральных каналов в технологии CWDM при использовании всей спектральной области от 1270 до 1610 нм в волокнах без водородного пика равно 18, число каналов в обычном одномодовом волокне еще меньше. Недостаточная масштабируемость систем CWDM может

быть преодолена внедрением гибридной технологии: DWDM поверх CWDM. Одну из возможных реализаций такой гибридной технологии иллюстрирует *рис. 8.6.*

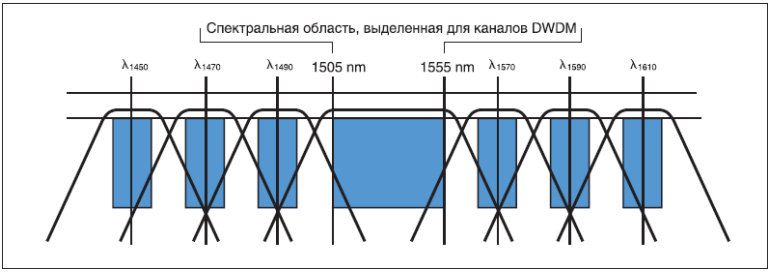

Рис.8.6. Реализация гибридной технологии DWDM и CWDM

Оптические мультиплексоры/демультиплексоры WDM

Оптические мультиплексоры/демультиплексоры являются центральными элементами WDM-систем. Они выполняют функцию объединения/разделения в пространстве спектральных каналов и фактически осуществляют пассивную маршрутизацию по длинам волн. Существуют различные оптические методы объединения и разделения нескольких каналов в одном волокне. Для разных видов WDM целесообразно применять методы, наиболее подходящие для них.

Двухволновые мультиплексоры.

Двухволновые WDM-мультиплексоры производятся с использованием хорошо зарекомендовавшей себя технологии сплавной биконической вытяжки (FBT), позволяющей достичь низких вносимых потерь одновременно с высокой степенью изоляции каналов в широком диапазоне температур.

Мультиплексоры на основе оптических фильтров

В мультиплексорах и демультиплексорах DWDM и CWDM могут быть использованы оптические узкополосные фильтры, каждый из которых выделяет из составного полихроматического светового пучка (или добавляет в него) один монохроматический пучок с определенной длиной волны. Применяются два типа узкополосных фильтров: тонкопленочные фильтры и фильтры на основе волоконных брэгговских решеток.

Тонкопленочный фильтр состоит из нескольких слоев прозрачного диэлектрического материала с различными показателями

преломления, нанесенных последовательно друг за другом на оптическую подложку.

Волоконная брэгговская решетка – это отрезок волокна с пространственной периодической структурой. Пространственная периодическая структура является объемной дифракционной решеткой – брэгговской решеткой, которая отражает свет в некотором диапазоне длин волн и пропускает свет всех остальных длин волн.

Мультиплексоры на основе дифракционных решеток

Дифракционные решетки отражают световой пучок некоторой длины волны под таким углом в плоскости падения, для которого разность набегов фаз от соседних элементов решетки равна 2π. Величина этого угла зависит от длины волны. Необходимость совмещения волоконных элементов с объемными делает устройства на основе дифракционных решеток дорогими и сложными в производстве. Однако вносимые ими потери практически не зависят от числа каналов, что делает эту технологию одной из наиболее привлекательных для использования в системах с большим числом каналов.

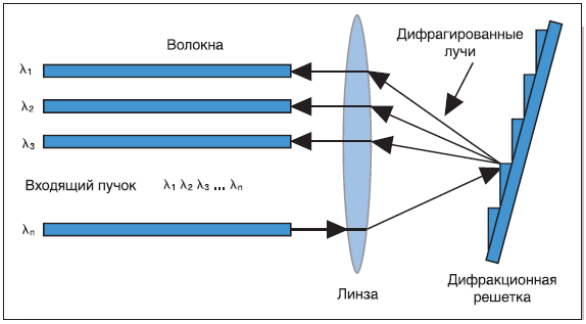

Рис.8.7. Оптическая схема демультиплексора DWDM на основе отражательной объемной дифракционной решетки

Упростить производство мультиплексоров с дифракционными решетками позволяет использование технологии интегральной оптики. Интегральная оптика успешно применяется для создания решеток на основе массива планарных волноводов различной длины AWG (Arrayed Waveguide Gratings). Принцип действия фазовой решетки состоит в том, что свет проходит через группу волноводов разной длины. AWG – это, по существу, эшелон Майкельсона в интегрально-оптическом исполнении.

Лекция 9. Оптические передатчики

Функции оптических передатчиков. Источники излучения в передатчиках с прямой модуляцией. Светодиоды в волоконно-оптических системах связи. Физический механизм работы лазеров.

Функции оптических передатчиков

Оптические передатчики (трансиверы), применяемые в волоконно-оптических системах связи, предназначены для преобразования электрических сигналов в оптические. С этой целью выходное излучение оптического источника модулируется в соответствии с входными электрическими сигналами, поступающими из передающей системы.

По характеру модуляции оптические передатчики делятся на передатчики с прямой (внутренней) и внешней модуляцией. В оптических передатчиках с прямой модуляцией мощность излучения источника света модулируется внешним электрическим током питания. В цифровых системах связи на основе передатчиков с прямой модуляцией используется простейший оптический формат передачи данных, при котором логическому нулю соответствует выключенное состояние источника излучения, а логической единице – включенное.

При скоростях передачи 10 Гбит/с и выше используются передатчики с внешней модуляцией. Источниками излучения в таких передатчиках, как правило, являются узкополосные одномодовые непрерывные полупроводниковые лазеры. Непрерывное оптическое излучение модулируется внешним модулятором, что обеспечивает формирование оптического сигнала с минимальной спектральной шириной. Кроме того, применение внешней модуляции позволяет использовать более сложные форматы модуляции и применять поляризационное разделение сигналов. Передатчики с внешней модуляцией используются в системах дальней связи, в которых требования к качеству оптического сигнала особенно высоки. Они позволяют передавать сигналы со скоростью в десятки гигабит в секунду на тысячи километров (с использованием оптических усилителей).

Максимальная скорость передачи информации, ограниченная быстродействием модулятора, составляет 40 Гбит/с. Для увеличения скорости передачи по одному каналу свыше 40 Гбит/с была предложена техника оптического временного уплотнения (OTDM). Такие системы работают со специальными источниками излучения –

лазерами с синхронизацией мод, которые генерируют непрерывную последовательность ультракоротких импульсов (УКИ) света.

Источники излучения в передатчиках с прямой модуляцией

Источниками излучения в оптических передатчиках с прямой модуляцией являются полупроводниковые светоизлучающие диоды (светодиоды) или лазеры. Передатчики на основе светодиодов используются совместно с многомодовым волокном в низкоскоростных системах передачи информации на короткие расстояния, но постепенно вытесняются лазерными передатчиками. Основными недостатками светодиодов являются малая скорость передачи информации, малая выходная мощность, широкая полоса спектра и большая расходимость излучения.

С другой стороны, светодиоды – более дешевые и неприхотливые приборы, которые вполне подходят для ряда систем небольшой протяженности или средней протяженности, но с малой скоростью передачи информации (менее 1 Гб/с). Поэтому они широко используются в замкнутых системах видеонаблюдения, в локальных вычислительных сетях (ЛВС), в измерительных и других сетях, построенных на основе многомодового оптического волокна. Применение светодиодов в аппаратуре связи позволяет существенно удешевить приемопередающее оборудование, что и является причиной использования кабелей с многомодовым волокном при строительстве ЛВС.

В городских сетях связи и системах дальней связи в качестве источников излучения используются полупроводниковые лазеры, обеспечивающие существенно большую вводимую в одномодовое волокно мощность, максимальную скорость передачи информации и обладающие существенно более узким спектром излучения по сравнению со светодиодами.

В системах связи со скоростью менее *2,5 Гбит/с* используются простейшие лазеры с резонатором Фабри – Перо и прямой модуляцией. При скоростях передачи информации $B = 2{,}5$ *Гбит/с* и выше необходимо использовать лазеры с распределенной обратной связью (РОС-лазеры) или лазеры с распределенными брэгговскими отражателями (РБО), в которых обеспечивается эффективная селекция мод и сужение спектра излучения. При скорости 2,5 Гбит/с допустимо использование прямой модуляции РОС-лазера.

В будущих сетях связи возможно широкое использование лазеров с вертикальным резонатором. Важнейшее потенциальное достоинство таких лазеров заключается в возможности массового производства и

тестирования (на одном полупроводниковом кристалле может быть изготовлено несколько сот лазеров одновременно), что, как ожидается, приведет к значительному снижению их стоимости.

Главное преимущество лазеров с прямой модуляцией – экономическое, т.к. такие устройства намного дешевле лазеров с внешней модуляцией. Главный недостаток – наличие паразитной частотной модуляции (ЧМ), или чирпа (Chirp). Чирп приводит к расширению спектра излучения и, как правило, к сокращению дальности широкополосной передачи информации.

Светодиоды в волоконно-оптических системах связи

Светодиоды применяются в системах волоконно-оптической связи, передающих данные на сравнительно короткие расстояния с низкими и средними скоростями. Поскольку время жизни спонтанного излучения в сильно возбужденных полупроводниках составляет около 1 нс, максимально достижимые скорости передачи данных в системах со светодиодами ограничены значением 1 Гбит/с. Поэтому в волоконно-оптических системах на основе светодиодов нельзя получить скорости передачи данных в несколько Гбит/с. Однако скорости передачи данных в несколько сотен Мбит/с обычно удовлетворяют требованиям большинства систем локальной связи.

Для получения максимальной эффективности ввода излучения светодиода в оптическое волокно область излучения светодиодов должна быть намного меньше диаметра сердцевины волокна. Для работы с многомодовыми волокнами, как правило, используются светодиоды с круговыми областями излучения с диаметром 20…50 мкм. Диаметр сердцевины многомодовых кварцевых волокон обычно равен 50…100 мкм.

Полимерные световоды могут иметь диаметры больше 1 мм, поэтому для работы с ними могут использоваться светодиоды с большей областью излучения.

Рассмотрим светодиод со временем нарастания сигнала τ_r. При подаче на вход светодиода ступенчатого импульса тока мощность выходного оптического излучения увеличивается по закону:

$$P_{out}(t) = P_0[1 - \exp(-t/\tau_r)] \,. \qquad (9.1)$$

Можно определить частотную *передаточную функцию светодиода по мощности* как зависимость глубины модуляции мощности от частоты при заданной глубине модуляции M тока:

$$P_{out}(\omega) = P_0 M (1 - i\omega\tau_r)^{-1} \,. \qquad (9.2)$$

Частота, при которой абсолютное значение этого выражения уменьшится в два раза, т.е. на 3 дБ, называется шириной полосы модуляции по уровню 3 дБ.

Максимальная частота модуляции светодиода определяется временем жизни неосновных носителей. Поэтому уменьшение времени их жизни за счет либо сильного легирования активной области, либо специального введения глубоких примесных центров, приводит к увеличению максимальной частоты модуляции светодиода. Введение глубоких примесных центров имеет двойной эффект. С одной стороны, наличие этих центров приводит к сокращению времени жизни неосновных носителей, тем самым увеличивая частоту по уровню 3 дБ. Но, с другой стороны, ведет к снижению интенсивности излучения и росту нагрева структуры светодиода. Высокие концентрации легирующих примесей в активной области также сокращают время жизни неосновных носителей и часто, но не всегда, ведут к уменьшению квантового выхода излучения диода.

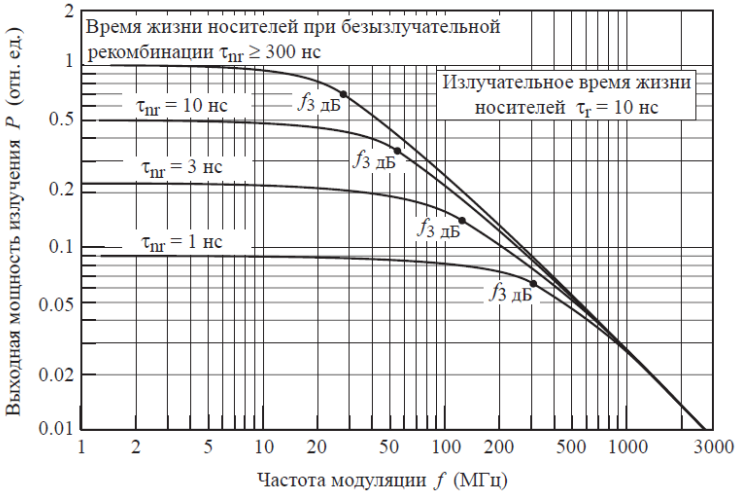

Рис. 9.1. Теоретические передаточные функции светодиода при различных значениях безизлучательного времени жизни (излучательное время жизни 10 нс).

Проанализируем влияние снижения времени жизни неосновных носителей на глубину модуляции и ширину полосы модуляции. Из

уравнения (9.2) можно получить следующее выражение для полосы модуляции по уровню 3 дБ:

$$f_{M0,5} = \sqrt{3}(2\pi\tau_{\Sigma})^{-1}, \text{ где } \tau_{\Sigma}^{-1} = \tau_r^{-1} + \tau_{nr}^{-1}. \quad (9.2)$$

Внутренний квантовый выход излучения светодиода определяется выражением: $\eta_{\text{int}} = \tau_{nr}/(\tau_r + \tau_{nr})$. В случае малого времени жизни носителей при безызлучательной рекомбинации: $f_{M0,5} \approx \tau_{nr}^{-1}, \eta_{\text{int}} \approx \tau_{nr}$. Отсюда видно, что хотя снижение времени безызлучательной релаксации, например за счет введения глубоких примесных центров, позволяет расширить полосу рабочих частот светодиода, произведение мощности модулированного излучения на полосу пропускания устройства при этом не улучшается. На рис.9.1 показано, как связаны между собой полоса модуляции по уровню 3 дБ, мощность модулированного излучения и время жизни носителей при излучательной и безызлучательной рекомбинации. Для вычисления частоты по уровню 3 дБ и значений мощности использовались уравнения, приведенные выше. Как видно из графика, построенном в логарифмическом масштабе, на частотах больших $f_{M0,5}$ происходит линейное снижение мощности модулированного сигнала. В линейных единицах (мВт) это означает экспоненциальное уменьшение мощности модулированного сигнала.

Пока методы снижения времени жизни неосновных носителей без уменьшения квантового выхода излучения светодиодов не найдены. Если это удастся сделать, будет достигнут реальный прогресс в увеличении быстродействия светодиодов.

Физический механизм работы лазеров

Физический механизм работы лазеров заключается в создании активной области (среды), в которой одновременно присутствуют носители зарядов двух типов: электроны, находящиеся в зоне проводимости, и дырки, находящиеся в валентной зоне. Вынужденная рекомбинация электронно-дырочных пар под действием световой волны вызывает усиление света в этой области. Полупроводник такого типа называется дважды вырожденным и распределение носителей заряда по энергетическим зонам такого полупроводника показано на рис. 2.2.г (См. Лекцию 2).

Если создать положительную обратную связь (поместить активную область кристалла в резонатор), то при превышении током инжекции некоторого (порогового) значения $(I_{inj})_n$ произойдет самовозбуждение лазерной генерации. Этот тип полупроводниковых лазеров называют инжекционными.

Первые полупроводниковые лазеры были созданы в 1962 г. в *США* и *СССР* на основе *p-n*-перехода в интерметаллическом соединении - арсениде галлия *GaAs*. Этот активный полупроводниковый кристалл используется и в настоящее время.

В первых инжекционных лазерах использовались электронно-дырочные переходы между *p*- и *n*- областями, созданными в одном и том же кристалле. Для этого соответствующие части кристалла легировались различным образом. Иначе говоря, *p*- и *n*-области отличались друг от друга только видом и концентрациями содержавшихся в них донорных и акцепторных примесей. Такие кристаллы называют гомоструктурами, приготовленные в них *p-n*-переходы - гомопереходами, а полупроводниковые лазеры на их основе - гомолазерами.

Наиболее существенным недостатком гомолазеров является высокое значение пороговой плотности тока инжекции (накачки) при комнатных температурах. Его происхождение поясняет схема активного элемента лазера на *p-n*-переходе, приведенная на рис. 9.2.

Толщина активного слоя в кристалле с *p-n*-переходом (заштрихованная область) по порядку величины равна среднему расстоянию, на которое успевает продиффундировать электрон, инжектированный из *n*- в *p*-область (до его рекомбинации с "встретившейся по дороге" дыркой). В граничном слое происходит изменение величин диэлектрической восприимчивости и показателя преломления активного кристалла, связанное в основном с градиентами содержания примесей (в меньшей степени - с дисперсией восприимчивости в области, где возникает инверсия населенностей и усиление). Градиент этих изменений уменьшается по мере удаления от средней плоскости слоя.

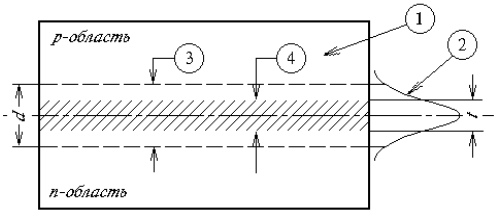

Рис.9.2. Схема гомолазера на p-n-переходе. 1) Кристалл с p-n-переходом. 2) Профиль интенсивности поля усиливаемой волны. 3) "Границы" объема, занимаемого волноводной модой. 4) Активный слой p-n-перехода.

Если электромагнитная волна распространяется в области, где имеет место изменение показателя преломления, то из-за волноводного эффекта волна оказывается локализованной в пространстве вблизи этой области. Это является причиной, по которой в резонаторе полупроводникового лазера возбуждаются диэлектрические волноводные моды. Интенсивность поля в этих модах быстро спадает по мере удаления от плоскостей, которые можно рассматривать как границы волновода (области локализации).

Однако толщина активного слоя (локализация инверсной населенности) *t*, в случае гомоперехода существенно меньше области локализации волноводной моды *d*. Так, например для *p-n*-перехода, приготовленного в кристалле *GaAs*, значение величины *t* составляет ~ 1 мкм, а величины *d* - ~ 2 - 5 мкм. Такое соотношение этих величин является следствием сравнительно небольшого перепада показателя преломления, возникающего в пограничной области (~ 0,1 - 1%).

Такое соотношение величин *d* и *t* является невыгодным по следующей причине. Т.к. $d > t$, то активный слой перекрывается лишь частью распространяющейся волны. Другая же ее часть оказывается за пределами этого слоя и при распространении не усиливается, а ослабляется. Очевидно, что если при неизменной толщине активного слоя *t* уменьшить область локализации волноводной моды *d*, то это приведет к снижению потерь и порогового значения плотности тока инжекции (накачки).

Если же $d < t$, то в процессе квантового усиления принимают участие не все активные частицы, т.к. часть из них оказывается вне основной части поля распространяющейся волны. Заметим, что в лазерах других типов в большинстве случаев имеет место именно такая ситуация. При таком соотношении *d* и *t* повышение эффективности действия лазера и снижение порогового значения плотности тока инжекции достигается уменьшением толщины активного слоя *p-n*-перехода *t* (сближении значений *d* и *t*). Оптимальным является соотношение $d = t$.

В первых полупроводниковых лазерах использовался *p-n*-гомопереход, и материалом для его создания был прямозонный кристалл арсенида галлия - *GaAs* .

Темп излучательных рекомбинационных переходов в прямозонных кристаллах более высок, чем в непрямозонных, т.е. в таких кристаллах, где рекомбинация носителей осуществляется путем непрямых переходов с участием фононов.

Радикальное решение, которое привело к одновременному уменьшению значений как d, так и t и обеспечило резкое снижение пороговой плотности тока инжекции, состояло в переходе к использованию при приготовлении p-n-переходов гетерогенных структур.

В гетерогенной структуре p-n-переход создается не в одном и том же кристалле, а на границе между различными кристаллам*и*, которые хотя и близки по структуре, но заметно отличаются по некоторым существенным (для получения перехода с малым значением пороговой плотности тока инжекции) параметрам. Такой переход называют гетеропереходом, а лазер на этом переходе - гетеролазером.

Первая проблема, которую необходимо было решить при создании p-n-гетеропереходов, состояла в поиске технологий, позволявших прочно соединить (сплавить) друг с другом два разнородных полупроводника. Это удается достигнуть далеко не для всякой пары кристаллов: при их сплавлении в пограничном слое могут возникать дефекты, делающие слой непрочным и склонным к разрушению.

Чтобы гетеропереход оказался близким к идеальному (не имеющему дефектов кристаллической решетки), приведенные в контакт кристаллы должны иметь однотипные решетки с совпадающими периодами (как показала практика, точность совпадения должна быть ~ 0,1%), одинаковые коэффициенты термического расширения.

Решить проблему удалось путем использования для создания гетероперехода пар родственных полупроводниковых кристаллических соединений, которые, отличаясь по составу и свойствам, имеют кристаллические решетки, которые чрезвычайно близки по своим параметрам.

Характерным примером таких родственных кристаллов являются арсенид галлия *AsGa* и арсенид алюминия *AsAl*. Оба кристалла являются соединениями элементов III и V групп периодической системы элементов и относятся к классу полупроводников $A^{III}B^{V}$.

В полупроводниковом кристалле, предназначенном для использования в гетеролазере, должен быть приготовлен, по крайней мере, один гетеропереход (а лучше - два или несколько). Полупроводниковый кристалл с несколькими гетеропереходами называют гетероструктурой.

На рис. 9.3 приведены схемы, дающие сравнительное представление о структурах и некоторых характеристиках гомолазера и гетеролазеров с *односторонним и двухсторонним ограничением активного слоя*.

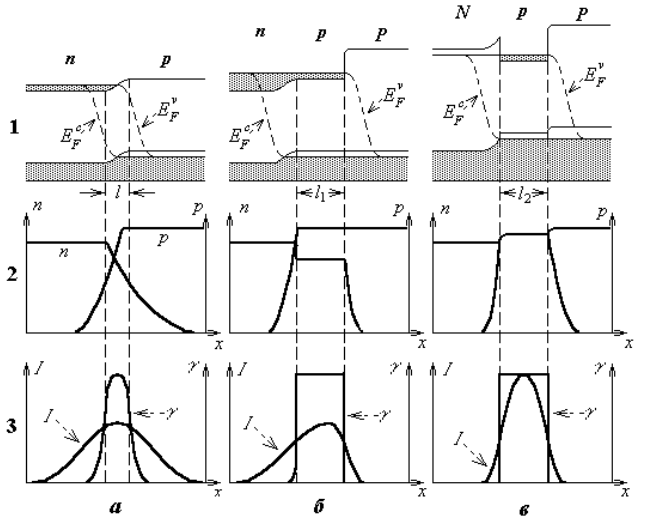

Рис. 9.3. Схемы и графики некоторых характеристик полупроводниковых лазеров на p-n-переходе разных типов: гомолазера (1а, 2а, 3а); гетеролазера с односторонним ограничением (1б, 2б, 3б); гетеролазера с двухсторонним ограничением (1в, 2в, 3в).
1) 1а, 1б, 1в - зонные схемы. Строчными буквами (p и n) обозначены области, образованные узкозонным, а прописными (P и N) - широкозонным полупроводником.
2) 2а, 2б, 2в - графики распределения по кристаллу концентраций электронов n и дырок p.
3) 3а, 3б, 3в - графики распределения по кристаллу интенсивности волны I в волноводной моде и коэффициента усиления в активном слое γ.

Первый случай - это уже рассмотренный ранее гомолазер. На схеме 1*а* *l* - это ширина области в переходном слое между *p*- и *n*-областями кристалла, в которой возникает инверсия населенностей (*l* = *t*) . Схемы 2*а*, 2*б* и 2*в* иллюстрируют распределение по кристаллу концентраций электронов и дырок в каждом из приведенных на данном рисунке случаев, а схемы 3*а,* 3*б* и 3*в* - распределение интенсивности электромагнитного поля *I* и коэффициента усиления γ.

Схемы 1*б*, 2*б* и 3*б* относятся ко *второму* случаю - гетеролазеру с односторонним ограничением. Здесь, как и в первом случае, имеется *p*-*n*-переход в кристалле с относительно менее широкой (среди используемой пары кристаллов) запрещенной зоной, и, кроме того, на расстоянии l_1 от этого *инжектирующего* перехода - гетеропереход на границе с кристаллом, имеющем более широкую запрещенную зону.

Первый кристалл будем называть *узкозонным* (*УЗ*) и обозначать образованные на его основе области в кристалле строчными буквами (*p*- и *n*-области), а второй - *широкозонным* (*ШЗ*), а для обозначения соответствующих областей в этом кристалле будем пользоваться прописными (заглавными) буквами (*P*- и *N*-области).

Во втором случае контакт между *УЗ* кристаллом, вблизи края которого расположен *p*-*n*-гомопереход, с *ШЗ* полупроводником создает в непосредственной близости от гомоперехода - *p*-*p*-гетеропереход. Роль этого перехода состоит, во-первых, *в пространственном ограничении* электронов, инжектированных *p*-*n*-гомопереходом в области между *p*-*n*-гомо- и *p*-*p*-гетеро- переходами (*электронное ограничение*), и во-вторых, - *в создании волноводного эффекта*, вызываемого перепадом значений показателя преломления на границе *УЗ* и *ШЗ* кристаллов (*оптическое ограничение*).

Электронное ограничение можно интерпретировать как отражение электронов, инжектированных *p*-*n*-гомопереходом из *n*-области в *p*-область *УЗ* полупроводника (где они являются неосновными носителями), от потенциального барьера, созданного *p*-*p*-гетеропереходом.

Цель оптического ограничения состоит в пространственной локализации электромагнитного поля усиливаемой волны. В случае гомоперехода вызывающий эту локализацию перепад показателя преломления $n_{пр}$ не создается специально (по этой причине имеющий место в этом случае волноводный эффект называют “случайным”), и сравнительно мал. В случае же гетероперехода перепад $n_{пр}$ происходит между областями, образованными на основе разных соединений, и его можно сделать значительно большим, чем в случае гомоперехода.

В *третьем* случае, показанном на рис. 9.3, активный слой кристалла гетеролазера ограничен двумя гетеропереходами. Здесь тонкий слой *УЗ* полупроводника шириной l_2, в котором образуется активная область, заключен между двумя *ШЗ* полупроводниками. Двухстороннее электронное и оптическое ограничение приводит к

тому, что ширины активной области t и волноводной моды d совпадают ($t = d = l_2$).

Среди трех рассмотренных выше полупроводниковых лазеров наилучшими параметрами обладает гетеролазер на основе двойной структуры.

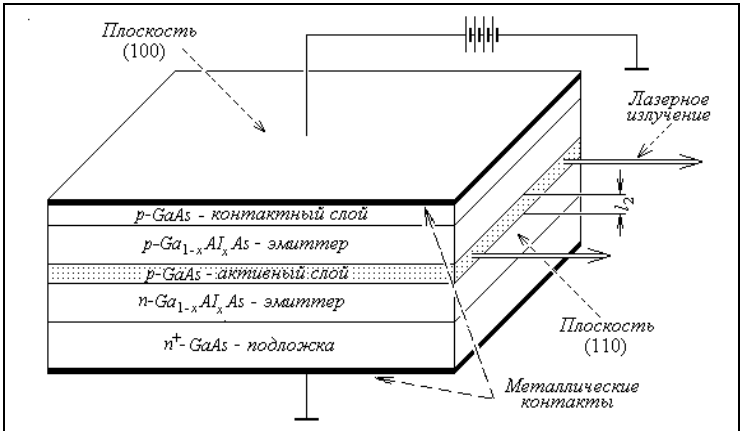

Рис. 9.4. Схематическое изображение гетеролазера на основе соединений GaAs - $Ga_{1-x}Al_xAs$. Резонатор образован плоскостями естественного скола кристалла (110), перпендикулярными плоскости активного слоя (100). l_2 – толщина активного слоя.

На рис. 9.4 схематически изображен гетеролазер с двухсторонним ограничением активного слоя. Им в этом случае является слой *УЗ* кристалла *p-GaAs*, расположенный между *ШЗ* слоями $n\text{-}Ga_{1-x}Al_xAs$ и $p\text{-}Ga_{1-x}Al_xAs$. Буквы *n* и *p* перед химической формулой образующего слой соединения обозначают тип его проводимости, созданный легированием. Из *ШЗ* слоев идет инжекция носителей в активный (*УЗ*) слой, почему их называют *эмиттерами*. Потенциальные барьеры на границах слоев-эмиттеров с активным слоем препятствуют обратной диффузии из него инжектированных носителей. Слои нанесены на *подложку* из $n^+ - GaAs$. Значок "+" около *n* говорит о *сильном легировании* материала подложки (по сравнению с нанесенным на нее слоем $n\text{-}Ga_{1-x}Al_xAs$). Поверх всех этих слоев нанесен *контактный* слой из *p-GaAs*. Внешние поверхности

контактного слоя и подложки металлизируются (золотятся), и к ним прикладывается внешнее напряжение.

Пусть x = 0,3. Для слоя *p-GaAs* ширина запрещенной зоны E_g равна 1,5 *эВ*, а для слоев $p-Ga_{0,7}Al_{0,3}As$ и $n-Ga_{0,7}Al_{0,3}As$ - 1,8 *эВ*. Эпитаксиальные методы приготовления (нанесения) слоев позволяют уменьшить толщину слоя *p-GaAs* до (0,1 - 0,4) *мкм*. Если в гомопереходе толщина активной области $t \sim 1$ *мкм*, то в данном случае благодаря ограничивающей роли гетеропереходов она снижается в несколько раз ($t \approx l_2$).

Значительно большей оказывается в данном случае и разница показателей преломления между активным слоем *p-GaAs* ($n_{пр} \approx 3{,}6$) и расположенными по обе стороны от него слоями $Ga_{0,7}Al_{0,3}As$ ($n_{пр} \approx$ 3,4). Если в гомолазере $\Delta n_{пр}$ составляет 0,1 - 1%, то в гетеролазере - ~5%. Поэтому оптический волновод, возникающий в гетеролазере, гораздо эффективнее удерживает усиливаемую волну. Величина d в этом случае мало отличается от t, и выходящие за пределы активного слоя “хвосты” волны, с которыми связаны потери излучения, значительно укорачиваются.

Реализуемое в гетеролазерах уменьшение величин t и d привело к резкому снижению порогового значения плотности тока инжекции и позволило создать лазеры этого типа, непрерывно генерирующие при комнатных температурах.

Оптические передатчики на основе лазеров с прямой модуляцией

В современных системах связи наиболее широко используются оптические передатчики на основе полупроводниковых лазеров с прямой модуляцией. Достоинством их по сравнению со светодиодами является существенно более узкий спектр излучения (особенно в одномодовом режиме генерации), более узкая диаграмма направленности излучения, позволяющая эффективно вводить его в одномодовые волокна, и более широкая полоса модуляции.

В городских сетях связи и системах дальней связи в качестве источников излучения обычно используются передатчики на основе простейших лазеров с резонатором Фабри – Перо и прямой модуляцией в сочетании с одномодовым волокном. В таких системах связи скорости передачи информации может достигать 2,5 Гбит/с.

В некоторых случаях экономически наиболее эффективно оказалось использование многомодового волокна в сочетании с лазерами с вертикальными резонаторами (VCSEL).

В системах дальней связи со скоростями 2,5 и 10 Гбит/с необходимо использовать одномодовые полопроводниковые лазеры с распределенной обратной связью (*DFB).*

Главное преимущество лазеров с прямой модуляцией – экономическое, т.к. такие устройства намного дешевле лазеров с внешней модуляцией. Главный недостаток – наличие паразитной частотной модуляции (ЧМ), или чирпа (Chirp). Чирп приводит к расширению спектра излучения и, как правило, к сокращению дальности широкополосной передачи информации.

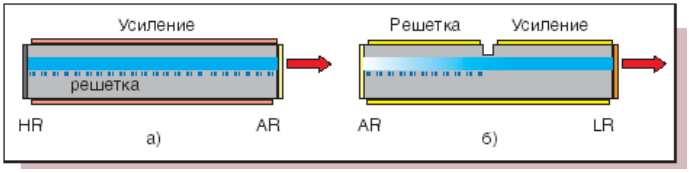

Рис. 9.5. Структурные схемы лазеров с периодическими структурами (решетками), используемыми для создания обратной связи: а) DFB_лазер; б) DBR_лазер. HR – зеркало с большим коэффициентом отражения; AR – просветляющее покрытие; LR – зеркало с низким коэффициентом отражения

Оптические передатчики на основе лазеров с внешней модуляцией

Источниками излучения в современных передатчиках протяженных телекоммуникационных систем передачи ВОСП являются непрерывные полупроводниковые лазеры. Для ослабления влияния хроматической дисперсии они должны работать в одномодовом одночастотном режиме, т.к. в этом случае достигается минимальная ширина спектра излучения.

Одним из решений этой задачи является использование лазеров с распределенной обратной связью. Вместо размещения зеркал на концах усиливающей области в ней самой создается периодическая решетка показателя преломления, как показано на рис. 9.5. Период решетки d подобран так, чтобы условие Брэгга выполнялось для отражения в обратном направлении. С учетом показателя преломления n условие Брэгга имеет вид $2nd = \lambda$. Условие отражения от периодической структуры выполняются для лучей обоих направлений. Таким образом, периодическая решетка создает обратную связь в обоих направлениях, распределенную по всей длине

лазера. Поскольку обратная связь, создаваемая периодической решеткой, является селективной, то в РОС-лазерах поддерживается одномодовый режим генерации. Другой перспективный тип полупроводниковых лазеров – лазеры с распределенными брэгговскими отражателями (РБО-лазеры), в которых также обеспечивается поддержание режима одночастотной одномодовой генерации.

Для создания информационного модулированного сигнала совместно с непрерывным полупроводниковым РОС-лазером используется внешний модулятор. Чаще всего применяются модуляторы на основе интерферометра Маха-Цендера (MZM) или электроабсорбционные модуляторы. На рис. 9.6 приведены блок-схемы передатчиков бинарных амплитудных форматов без возвращения к нулю (NRZ) и с возвращением к нулю (RZ) с различными рабочими циклами (50%; 33% и 67%).

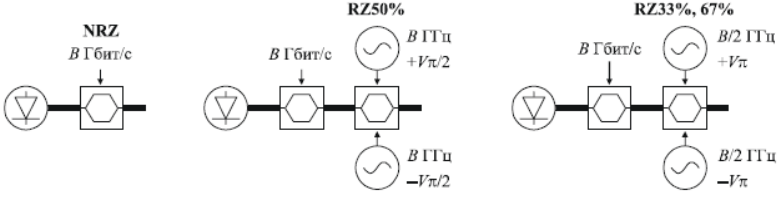

Рис. 9.6. Блок-схемы передатчиков бинарных амплитудных форматов (NRZ и RZ) с внешними модуляторами на основе интерферометров Маха-Цендера (MZM).

Комплексная амплитудная передаточная функция интерферометра Маха-Цандера имеет вид:

$$T_E(V_1,V_2) = \tfrac{1}{2}\{\exp[\,j\phi_1(V_1)] + \exp[\,j\phi_2(V_2) + j\psi]\} = \\ = \exp\{\,j[\phi_1(V_1) + \phi_2(V_2) + \psi]/2\}\cos\{[\phi_1(V_1) - \phi_2(V_2) - \psi]/2\} \qquad (9.3)$$

Коэффициент пропускания по мощности

$$T_P(V_1,V_2) = |T_E(V_1,V_2)|^2 = \cos^2\{[\phi_1(V_1) - \phi_2(V_2) - \psi]/2\} \qquad (9.4)$$

Обычно выполняются следующие условия:

$$\phi_1(V) = \phi_2(V) = \phi(V),\ \phi(V) = \kappa V,\ \psi = \kappa V_{Bias}. \qquad (9.5)$$

Тогда справедливы выражения ($\Delta V = V_1 - V_2, V_\Sigma = V_1 + V_2$):

$$T_E(V_1,V_2) = \exp\{\,j[\kappa V_\Sigma + \kappa V_{Bias}]/2\}\cos\{[\kappa\Delta V - \kappa V_{Bias}]/2\} \qquad (9.6)$$

$$T_P(V_1,V_2) = T_P(\Delta V) = \cos^2\{[\kappa\Delta V - \kappa V_{Bias}]/2\}. \qquad (9.7)$$

Лекция 10. Оптические приемники цифровых систем связи

Оптические цифровые приемники. Чувствительность приемников оптического излучения. Полупроводниковые фотодиоды. Источники шумов в цифровых оптических линиях связи. Тепловой шум. Квантовый предел чувствительности

Оптические цифровые приемники

Приемники оптического излучения (фотоприемники) в цифровых системах связи представляют собой сложные устройства, осуществляющие преобразование световых сигналов в электрические. Для этого световое излучение преобразуется в электрический ток, усиливается, а затем происходит восстановление переданного сообщения и формирование соответствующего этому сообщению электрического сигнала. Подавляющее большинство действующих оптических систем передачи информации используют двоичный (бинарный) код и простейшую амплитудную модуляцию с двумя значениями амплитуды сигнала. Приемники оптического излучения для таких систем имеют наиболее простую структуру.

В последнее время в научных лабораториях интенсивно исследуются различные новые форматы модуляции. Приемники для таких систем имеют более сложную структуру, но и в них составной частью присутствуют приемники бинарных амплитудно-модулированных сигналов.

Схема цифрового приемника амплитудно модулированного бинарного сигнала приведена на рис. 10.1.

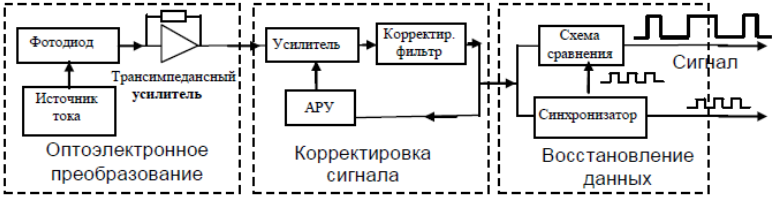

Рис.10.1. Схема приемник цифровой волоконно-оптической системы связи с амплитудной модуляцией и прямым детектированием.

В первом блоке цифрового фотоприемника оптический сигнал преобразуется в электрический сигнал с помощью фотодиода, установленного на входе в трансимпедансный усилитель.

Во втором блоке осуществляется усиление и коррекция АЧХ (амплитудно-частотной характеристики) усилителей и фотодиода. Система АРУ обеспечивает постоянство амплитуды сигнала на входе в регенератор. Фильтр на выходе линейного усилителя корректирует АЧХ так, чтобы добиться максимального отношения сигнал/шум при минимальной величине межсимвольной интерференции.

В третьем блоке восстанавливается исходная импульсная последовательность. Для принятия решения о том, какой символ передан (1 или 0) последовательность регенерируемых импульсов сравнивается с пороговым уровнем. Сравнение производится в середине тактового периода. Необходимый для этого синхросигнал формируется из тактовой последовательности регенерируемых импульсов. Схема сравнения восстанавливает первоначальную последовательность цифровых символов.

Полупроводниковые фотодиоды

Принцип работы фотодиода заключается в поглощении фотонов падающего излучения с образованием электрон-дырочной пары в полупроводнике. При наличии внешнего напряжения образовавшиеся пары носителей заряда создают электрический ток, называемый *фототоком*.

Важнейшей характеристикой фотодиодов является токовая чувствительность, равная отношению фототока I_p к поглощенной мощности света P_{in} и измеряемая в А/Вт.

$$S_{A/W} = \frac{I_p}{P_{in}} = \frac{e\eta}{h\upsilon_C}, \qquad (10.1)$$

где e - заряд электрона, η - квантовый выход люминесценции (отношение числа фотоэлектронов к числу падающих фотонов), υ_c - частота световой волны, h - постоянная планка.

Токовая чувствительность может быть явно выражена через длину волны света:

$$S_{A/W} = \frac{\eta\lambda}{1{,}24}\ [\text{А/Вт}]. \qquad (10.2)$$

Токовая чувствительность увеличивается с ростом длины волны, пока не достигнет красной границы материала, из которого изготовлен фотоприемник, после чего она резко уменьшается. В современных системах связи в диапазоне 1310 и 1550 нм

используются pin-фотодиоды на основе InGaAs, токовая чувствительность которых на $\lambda \sim 1550$ нм равна примерно 1 А/Вт.

Типы фотодиодов. В оптических линиях передачи применяются два типа фотодиодов: pin диоды и лавинные фотодиоды (ЛФД). Они имеют малые размеры и хорошо стыкуются с оптическими волокнами и с полупроводниковыми микросхемами. Достоинство ЛФД в том, что за счет внутреннего усиления фототока чувствительность фотоприемников с ЛФД получается в среднем на 6 дБ выше, чем у фотоприемников с pin диодами. А недостаток в том, что ЛФД работают при высоких напряжениях смещения (80 – 400 В) и их необходимо термостабилизировать (из-за сильной зависимости коэффициента лавинного усиления от температуры). В тоже время pin диод питается от того же источника, что и полупроводниковые микросхемы не требует термостабилизации. Время наработки на отказ у него примерно в 10 раз больше чем у ЛФД.

Таблица 10.1. Параметры pin диодов

Параметр	Размерность	Si	Ge	InGaAs
Длина волны	мкм	0.4 – 1.1	0.8 – 1.8	1.0 – 1.7
Токовая чувствительность	А/Вт	0.4 – 0.6	0.5 – 0.7	0.6 – 0.9
Квантовая эффективность	%	75 - 90	50 - 55	60 – 70
Темновой ток	нА	1 - 10	50 - 500	1 – 20
Ширина полосы	ГГц	0.3 – 0.6	0.5 - 3	1 - 10
Напряжение смещения	В	50 - 100	6 - 10	5 - 6

В линиях передачи с оптическими усилителями фотодиоды с внутренним усилением (ЛФД) теряют свое основное преимущество, так как чувствительность фотоприемников с оптическим предусилителем на входе примерно на 6 дБ выше, чем у фотоприемников с ЛФД. Причем при достаточно большом коэффициенте усиления оптического предусилителя чувствительность фотоприемника уже не зависит от типа фотодиода. Поэтому в линиях с оптическими усилителями применяются pin диоды, так как они проще в эксплуатации. Рабочая длина волны pin диода определяется шириной запрещенной зоны среднего i– слоя, где в основном поглощаются фотоны. Германиевые фотодиоды применяются на длине волны 1.3 мкм, а в высокоскоростных линиях передачи на

длине волны 1.55 мкм фотодиоды из InGaAs, так как они обладают более высоким быстродействием.

Структура pin диода на основе InGaAs приведена на рис. 2а. Из рис. 2а видно, что n и p слои состоят из InP, а i – слой из InGaAs. Ширина запрещенной зоны у InP больше, чем у InGaAs, то длина волны красной границы фотоэффекта у InP меньше. Так как слои из InP поглощают свет на $\lambda < 0.92$ мкм, а i – слой из InGaAs на $\lambda < 1.65$ мкм, то диапазон рабочих длин волн pin диода лежит в пределах от 1..1.7 мкм (табл. 1).

Как видно из таблицы 10.1 pin диод на основе InGaAs обладает заметно более высоким быстродействием, чем германиевые и кремниевые pin диоды. Его быстродействие ограничивается только временем дрейфа носителей в i – слое (~ 10 пс). Обусловлено это тем, что в pin диоде с i - слоем на основе InGaAs соседние слои p и n изготовлены из другого материала (InP). В такой структуре свет на рабочей длине волны поглощается только в i – слое и, т.о., исключается медленная диффузионная компонента в фототоке. В германиевых и кремниевых pin диодах p/n переход изготовлен из одного материала, и свет на рабочей длине волны поглощается также и в n и p слоях, что и приводит к появлению диффузионной компоненты в фототоке.

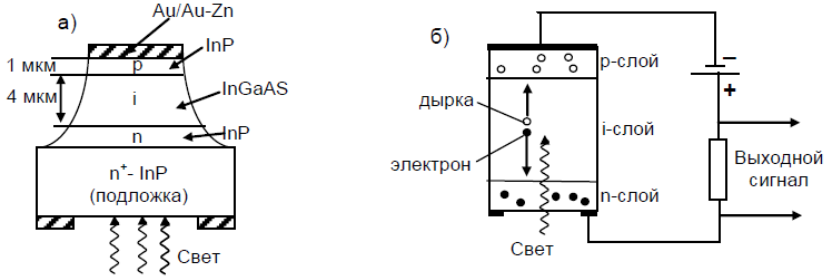

Рис. 10.2. (а) Структура pin диода на основе InGaAs (б) Поглощение света в i– слое с образованием электронно-дырочной пары

Механизм детектирования света поясняется на рис. 2б. На pin диод подается обратное напряжение (плюс со стороны n –слоя). Поэтому электроны уходят в n – слой, дырки в p – слой, а i – слой обедняется. Распределение электрического поля в этих слоях получается как в конденсаторе, где пластинами служат p и n слои. Так

как электропроводность обедненного i – слоя много меньше чем у p и n слоев, то все электрическое поле сосредоточено в i – слое.

Каждый фотон, поглощенный в i – слое, вызывает переход электрона из валентной зоны в зону проводимости, и приводит к рождению пары свободных носителей: электрона и дырки (рис. 10.2б). Электрическое поле разделяет и разгоняет появившиеся в результате фотоэффекта свободные носители и вызывает фототок в цепи смещения пропорциональный падающей на pin диод оптической мощности: I = S P, коэффициент пропорциональности, называется токовой чувствительностью и определяется выражением 10.2.

Чувствительность приемников оптического излучения

Важнейшей рабочей характеристикой действующей системы передачи информации, определяющей качество связи, является коэффициент ошибок. Его значение равно отношению числа ошибочно интерпретированных символов к общему числу переданных символов. Причина возникновения ошибок – наличие шумов. Действительно, в реальных системах связи значения фототока, соответствующие и 1, и 0, флуктуируют во времени из-за наличия шумов. Такие временные флуктуации тока могут привести к ошибочной интерпретации информационного символа.

Природу возникновения ошибок в двоичных цифровых системах связи с амплитудной модуляцией поясняет рис. 10.3.

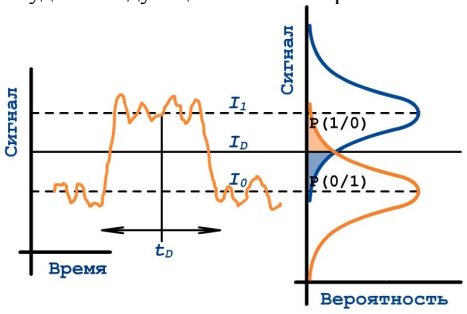

Рис. 10.3. Электрический информационный сигнал с шумом на входе схемы сравнения, уровень нуля I_0, уровень единицы I_1, уровень сравнения I_D, длительность такта t_D (слева) и распределения вероятностей измеренных значений тока сигнала для 1 и 0 (справа). Закрашенные области показывают вероятности ошибок: P(1/0) – вероятность интерпретации 0 как 1; P(0/1) – вероятность интерпретации 1 как 0

Из-за наличия шумов измеренное значение тока отличается от его точного значения. Разброс измеренных значений тока при передаче логической 1 и 0 описывается соответствующими функциями $F_1(I)$ и $F_0(I)$ распределения вероятностей. На рис. 10.3, справа, графики функций $F_1(I)$ и $F_0(I)$ показаны соответственно верхней и нижней кривыми. Как видно из рисунка, графики этих функций пересекают прямую, соответствующую уровню напряжения сравнения I_D. Это означает, что существует некоторая, обычно весьма малая, но отличная от 0 вероятность неправильной интерпретации принятого сигнала. Вероятность P(1/0) ошибочной интерпретации 0 как 1 определяется площадью под частью функции распределения $F_0(I)$, отсекаемой уровнем тока сравнения I_D. Аналогично вероятность P (0/1) ошибочной интерпретации 1 как 0 определяется площадью под частью функции распределения $F_1(I)$, отсекаемой уровнем тока сравнения I_D. При равной вероятности передачи 0 и 1 коэффициент ошибок определяется простым выражением

$K_{ОШ}$ = 1 / 2 (P (1/0) + P (0/1)).

В предположении гауссовского распределения шума с нулевыми средними значениями интенсивности и со среднеквадратичными отклонениями σ_1, σ_2 для 1 и 0 соответственно коэффициент ошибки определяется выражением:

$$K_{ОШ} = PE(Q), \qquad (10.3)$$

где функция $PE(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} dy \exp\left(-\frac{1}{2}(y)^2\right)$, а аргумент $Q = \frac{I_1 - I_2}{\sigma_1 + \sigma_2}$ – называется показателем качества принимаемого сигнала или его Q-фактором.

Для нормальной работы цифровой системы связи требуется, чтобы коэффициент ошибок $K_{ОШ}$ не превышал некоторого заданного значения. Обычно $K_{ОШ}$ уменьшается при увеличении амплитуды полезного сигнала и увеличивается при его уменьшении (обратная зависимость может наблюдаться при перегрузке фотоприемника). Минимальное значение средней мощности оптического излучения, необходимое для передачи сигналов с заданным коэффициентом ошибок, называется ***чувствительностью оптического приемника***.

Следует помнить, что заданный уровень может быть разным в зависимости от стандарта или требований оператора связи. Обычно он лежит в пределах от 10^{-9} до 10^{-12}.

Коэффициенту ошибок, равному 10^{-12} соответствует значение Q-фактора, равное 7.

Чувствительность может выражаться в линейных единицах, производных от ватта (нВт, мкВт) или в логарифмических – децибелах по отношению к милливатту (дБм).

Реальная чувствительность приемников определяется многими факторами: нормируемым значением коэффициента ошибок, формой импульса, скоростью передачи информации, шириной полосы приемника и шумами оптического излучения. Поэтому практически в спецификациях чувствительность приемника задается только для вполне определенного передатчика, скорости передачи двоичных сигналов и их формы.

С увеличением скорости передачи информации чувствительность ухудшается (т.е. возрастает) в линейных единицах приблизительно пропорционально скорости B [бит/с].

Чувствительность современных цифровых высокоскоростных приемников на основе pin-фотодиодов определяется тепловыми шумами электрического предусилителя (обычно в качестве предусилителя используется трансимпедансный усилитель).

В отсутствии тепловых шумов чувствительность фотоприемника определяется квантовыми свойствами светового излучения и называется квантовым пределом чувствительности.

Зависимость коэффициента ошибок от Q-фактора

При Q > 3 коэффициент ошибок $K_{ОШ}$ с точностью ~ 1 % аппроксимируется первым членом в асимптотическом разложении функции $\mathrm{erfc}(Q/\sqrt{2})$:

$$K_{ОШ} \cong (1/Q\sqrt{2\pi})\exp(-Q^2/2) \qquad (10.4)$$

Зависимость $K_{ОШ}$ от Q приведена на рис. 10.4.

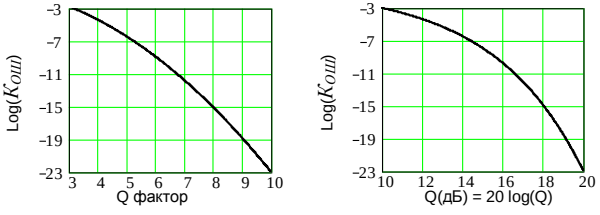

Рис. 10.4. Зависимость коэффициента ошибок ($K_{ОШ}$) от Q фактора

Из этого рисунка, в частности, видно, что BER ≅ 10^{-9} при Q = 6, и BER ≅ 10^{-12} при Q = 7. Значения Q – фактора часто приводят в логарифмических единицах: Q(дБ) = 20 log Q. Q(дБ) ≅ 15.6 дБ при Q = 6 и а Q(дБ) ≅ 16.8 дБ при Q = 7. Поскольку в общем случае $K_{ОШ}$ однозначно определяется величиной Q, то Q – фактор часто используется для характеристики качества передаваемого сигнала.

Источники шумов в цифровых оптических линиях связи

В оптическом цифровом приемнике происходит преобразование оптического сигнала в электрический, однако наряду с фототоком, пропорциональным мощности оптического сигнала, возникает также случайный шумовой ток. Существуют три фундаментальных механизма (причины) возникновения шумов фотоприемника.

1. Тепловой шум
2. Дробовый (квантовый) шум
3. Шум усиленного спонтанного излучения

Тепловой шум является следствием случайного движения электронов и существует всегда при температуре, отличной от абсолютного нуля. Он является доминирующим при работе систем связи без промежуточных оптических усилителей в нормальных условиях.

Дробовый шум является проявлением дискретной природы света (если использовать квантовые представления) или дискретного характера фототока (при использовании классического описания процесса фотодетектирования). Этот вид шума определяет фундаментальный предел чувствительности фотоприемников, если другие виды шумов устранены.

Шум усиленного спонтанного излучения, как и следует из его названия, вызывается спонтанным излучением в оптических усилителях и его последующим усилением.

Тепловой шум

Тепловой шум на сопротивлении R при температуре T можно описать как белый шум со средним значением равным нулю. При ширине полосы усилителя B_e (электрическая полоса) среднеквадратичное отклонение теплового тока имеет значение

$$\sigma_{therm}^2 = (4k_B T / R)B_e, \qquad (10.5)$$

где $k_B = 1{,}38 \cdot 10^{-23}$ Дж/К - постоянная Больцмана. Ширина полосы усилителя оптимизируется под конкретную схему модуляции и

обычно лежит в диапазоне от B до $B/2$ Гц, где B - битовая скорость передачи информации.

В предусилителе добавляются шумы других элементов. Обычно это явление учитывается введением некоторого коэффициента, называемого шум-фактором NF_e предусилителя. С учетом этого вклад теплового шума определяется среднеквадратичным значением тока

$$\sigma_{therm}^2 = (4k_B T / R)(NF_e)B_e \equiv I_t^2 B_e \qquad (10.6)$$

где введен параметр I_t, используемый для представления шумовых характеристик фотодиода. Типичное значение I_t имеет порядок единиц $nA/\sqrt{Гц}$. Величина шум-фактора обычно лежит в пределах от 2 до 3 (или в логарифмических единицах 3-6 дБ).

Подставив типичные значения параметров для нормальных условий работы ($B_e = B/2$Гц, NF_e=2 (3 дБ), R=100 Ом, T=300 К) получим:

$$\sigma_{therm}^2 = (4k_B T / R)(NF_e)B_e = 1{,}656 \cdot 10^{-22} B \quad [A^2] \qquad (10.7)$$

Оценим чувствительность фотодиода если наибольший вклад в шум вносит тепловой механизм (другими источниками шумов пренебрежем). Для этого по заданному значению коэффициента ошибки $K_{ОШ}$ найдем величину Q-фактора.

$$Q = \frac{I_1 - I_2}{\sigma_1 + \sigma_2} = (PE)^{-1}(K_{ОШ}) \qquad (10.8)$$

Считая $I_2 = 0$, а $\sigma_1 = \sigma_2 = \sigma_{therm}$ для тока I_1 получим выражение:

$$I_1 = (\sigma_1 + \sigma_2)Q = 2\sigma_{therm}Q \qquad (10.9)$$

Учитывая, что $P_{sens} = (1/2)P_1 = (1/2)I_1 / S_{AW}$ окончательно получим:

$$P_{sens} = (1/2)I_1 / S_{AW} = \frac{\sigma_{therm} Q}{S_{AW}} = \frac{1{,}3 \cdot 10^{-11} \cdot \sqrt{B} \cdot 7}{1{,}25} \text{ [Вт]}. \qquad (10.10)$$

Квантовый предел чувствительности

Дробовый шум становится доминирующим если удается устранить тепловой шум, например, при использовании приемников, охлаждаемых до гелиевых температур. Рассмотрим, какова может быть предельно достижимая чувствительность фотоприемника, в котором полностью устранен тепловой шум и отсутствуют шумы спонтанного излучения.

В этом случае приемник должен работать в режиме счета фотонов: если не зарегистрирован ни один фотон, то приемник считает, что принят 0, если зарегистрирован 1 или более фотонов, то принимается решение, что поступивший сигнал есть 1.

Рассматриваемый идеальный приемник регистрирует 0 без ошибок, а вот 1 может быть ошибочно принята 0 если не удастся зарегистрировать ни одного фотона.

Вероятность такой ошибки дается выражением:

$p(0/1) = \exp(-M)$, где M - среднее число фотонов на бит, представляющий 1.

$$K_{ОШ} = 1/2\, p(0/1) = 1/2 \exp(-M) \qquad (10.11)$$

Шумы спонтанного излучения

В линиях с оптическими усилителями обычно преобладают шумы, вызванные спонтанным излучением, которые принято называть шумами усиленного спонтанного излучения. Природа этих шумов – спонтанное оптическое излучение возбужденных частиц оптических усилителей.

Шум усиленного спонтанного излучения – это классический шум и он ведет себя аналогично полезному сигналу, т.е. ослабляется в линии и усиливается в усилителе. В каждом усилителе к усиленному классическому шуму от предыдущего каскада добавляется шум усиленного спонтанного излучения самого усилителя. Спектр усиленного спонтанного излучения намного шире спектра информационного сигнала, поэтому качество оптического сигнала характеризуют величиной, которую принято называть оптическим отношением сигнала к шуму (*OSNR*). *OSNR* равно отношению мощности полезного сигнала к мощности шума в некоторой стандартной (или специально оговоренной) спектральной оптической полосе. Поскольку измерения *OSNR* проводятся при помощи оптического анализатора спектра (OSA), то ширину полосы обозначим $\Delta\nu_{OSA}$.

При попадании на оптический приемник шум усиленного спонтанного излучения приводит к возникновению шумового тока фотодетектора, дисперсия которого описывается выражением:

$$\sigma^2 = \sigma^2_{с-сп} + \sigma^2_{сп-сп}, \qquad (10.12)$$

где $\sigma^2_{с-сп}$ – дисперсия шумов биений сигнала со спонтанным излучением, $\sigma^2_{сп-сп}$ – дисперсия шумов биений спектральных компонент спонтанного излучения между собой. Если ширина оптической полосы $\Delta\nu_D$ много больше электрической полосы Δf приемника можно получить следующие выражения для величин дисперсии этих двух компонент шума:

$$\sigma^2_{с-сп} = 2 I_C I_{сп} (\Delta f / \Delta\nu_D), \qquad (10.13)$$

$$\sigma^2_{сп-сп} = I^2_{сп} (\Delta f / \Delta\nu_D), \qquad (10.14)$$

где $I_c = S_{A/W} P_c$, $I_{cn} = S_{A/W} P_{cn}$, а P_c и P_{cn} – мощность сигнала и мощность шума на входе фотоприемника. Следует иметь в виду, что входящее в (10.13) выражение для мощности сигнала P_c есть некоторое усредненное за время $t_f = 1/\Delta f$ около моментов принятия решения значение мгновенной мощности сигнала.

Следует обратить внимание, что в формулах (10.13) и (10.14) величина I_{cn} пропорциональна ширине оптической полосы $\Delta\nu_D$ в силу того, что $P_{cn} = \rho_{cn}\Delta\nu_D$, где ρ_{cn} – спектральная плотность мощности шума УСИ. Следовательно, дисперсия шумов биений сигнала со спонтанным излучением не зависит от ширины оптической полосы $\Delta\nu_D$, а дисперсия шумов биений спектральных компонент спонтанного излучения между собой даже прямо пропорциональна $\Delta\nu_D$.

$$\sigma^2_{c-cn} = 2S^2_{A/W} P_C \rho_{cn} \Delta f\,, \tag{10.15}$$

$$\sigma^2_{cm-cn} = S^2_{A/W} \rho^2_{cn} (\Delta f \cdot \Delta\nu_D)\,. \tag{10.16}$$

Плотность мощности усиленного спонтанного излучения экспериментально определяется по измеренному значению мощности сигнала и OSNR:

$$\rho_{cn} = \frac{\overline{P}_C}{(OSNR \cdot \Delta\nu_{OSA})}, \tag{10.17}$$

где $\overline{P}_C$ – среднее по времени значение мощности сигнала.

С учетом (10.17) выражения для двух компонент шумов биений примут вид:

$$\sigma^2_{c-cn} = \frac{2S^2_{A/W} P_C \overline{P}_C \Delta f}{OSNR \cdot \Delta\nu_{OSA}}, \tag{10.18}$$

$$\sigma^2_{cm-cn} = \frac{S^2_{A/W} \overline{P}^2_C (\Delta f \cdot \Delta\nu_D)}{(OSNR \cdot \Delta\nu_{OSA})^2}. \tag{10.19}$$

В современных высокоскоростных системах связи вклад шума биений сигнала со спонтанным излучением является, как правило, определяющим. В этом случае можно оценить требуемую величину OSNR, которая обеспечивает заданный уровень коэффициента ошибок $K_{ош}$. Предполагаем, что формат передачи информации NRZ (без возвращения к нулю). В этом случае $P_C = 2\overline{P}_C$ для 1 и $P_C = 0$ для 0. Найдем величину $Q = \dfrac{I_1 - I_0}{\sigma_1 + \sigma_0}$, которая при сделанных предположениях ($I_0 = 0, \sigma_0 = 0$) определяется выражением:

$$Q = \frac{I_1}{\sigma_1} = S_{A/W} P \Big/ \sqrt{\frac{2S_{A/W}^2 P_C \bar{P}_C \Delta f}{OSNR \cdot \Delta \nu_{OSA}}} = \sqrt{\frac{OSNR \cdot \Delta \nu_{OSA}}{\Delta f}} \,. \qquad (10.20)$$

По заданному уровню $K_{ош}$ определяется требуемое значение $Q \equiv Q_T$, зная который можно найти требуемое значение $OSNR \equiv OSNR_T$:

$$OSNR_T = Q_T^2 \frac{\Delta f}{\Delta \nu_{OSA}} \,. \qquad (10.21)$$

Поскольку для $K_{ОШ} = 10^{-9}$ величина $Q_T = 6$, то для требуемого $OSNR_T$ получаем следующее приближенное выражение:

$$OSNR_T = 36 \frac{\Delta f}{\Delta \nu_{OSA}} \,. \qquad (10.22)$$

Если используются коды с коррекцией ошибок (FEC), то величина $K_{ОШ}$ (до коррекции) может быть снижена до уровня 10^{-3}, которому соотвествует величина $Q_T = 3$. В этом случае для требуемого $OSNR_T$ получаем следующее приближенное выражение:

$$OSNR_T = 9 \frac{\Delta f}{\Delta \nu_{OSA}} \,. \qquad (10.23)$$

Следует оценить правомерность пренебрежения шумами биений спонтанного излучения со спонтанным излучением. Для этого найдем отношение $\sigma_{с-сп}^2$ к $\sigma_{ст-сп}^2$:

$$\frac{\sigma_{с-сп}^2}{\sigma_{ст-сп}^2} = \frac{4 \cdot OSNR \cdot \Delta \nu_{OSA}}{\Delta \nu_D} \,. \qquad (10.24)$$

Если отношение сигнал/шум определяется выражением (10.23), то отношение вкладов двух компонент шума равно:

$$\frac{\sigma_{с-сп}^2}{\sigma_{ст-сп}^2} = \frac{36 \cdot \Delta f}{\Delta \nu_D} \,. \qquad (10.25)$$

Таким образом, поскольку ширина полосы оптического фильтра может на порядок превышать величину электрического фильтра, то в системах связи с использованием коррекции ошибок не всегда правомерно пренебрежение шумом биений спонтанного излучения со спонтанным излучением.

Литература

1. Наний О.Е. Основы цифровых волоконно-оптических систем связи. Lightwave Russian Edition, 2003, №1, с. 48 – 52.
2. Листвин А. В., Листвин В. Н., Швырков Д. В., Оптические волокна для линий связи, М.: ЛЕСАРарт, 2003.
3. Шуберт Ф.Е. Светодиоды. М.: «Физматлит», 2008.
4. Пихтин А.Н. Оптическая и квантовая электроника. М.: «Высшая школа», 2001.
5. Корниенко Л.С., Наний О.Е. Физика лазеров. М.: Издательство МГУ, 1996.
6. Гауэр Дж. Оптические системы связи. М.: «Радио и связь», 1989.
7. Козане А., Флере Ж., Мэтр Г., Руссо М. Оптика и связь. М.: «Мир», 1984.
8. Волоконная оптика, сборник статей. М., ВиКо, 2002.
9. Убайдуллаев Р.Р. Протяженные ВОЛС на основе EDFA. Lightwave Russian Edition, 2003, №1, с. 22 – 28.
10. Скляр Б., Цифровая связь. М: «Вильямс», 2004.
11. Наний О.Е. Оптические передатчики, Lightwave Russian Edition, 2003, №2, с. 48-51.
12. Наний О.Е. Приемники цифровых волоконно-оптических систем связи, Lightwave Russian Edition, 2004, № 1, с. 50-51.
13. Фриман Р. Волоконно-оптические системы связи. М.: «Техносфера», 2003.
14. Величко М., Наний О., Сусьян А. Новые форматы модуляции в оптических системах связи, Lightwave Russian Edition, 2005, № 4, с. 21-30.
15. Agrawal G. Lightwave Technology Telecommunication Systems. Wiley, New Jersey, 2005.